KYBERNETIKA — VOLUME 33 (1997), NUMBER 2, PAGES 133 - 159

ASYMPTOTIC RESULTS IN PARAMETER ESTIMATION FOR GIBBS RANDOM FIELDS[1]

MARTIN JANŽURA

Both the maximum likelihood estimate and a class of the maximum pseudo-likelihood estimates for parameters of Gibbs random fields are introduced, and their asymptotic properties, namely the consistency, the asymptotic normality, and the asymptotic efficiency, are studied, as well as the interrelations between the particular estimators and their respective properties.

1. INTRODUCTION

The statistical inference for Gibbs distributions has been recently widely studied because of its relevance for image processing and spatial statistics. The Gibbs distributions were originally used in frame of statistical physics to describe the equilibrium states of large systems. For the "statistical" purposes they seem to be rather justified since they obey both the physical experience and the intuitive mathematical assumption of maximum entropy and local dependence structure. They can be also understood as an infinite-dimensional generalization of the usual exponential family of distributions or the log-linear models for contingency tables data.

A natural parametrization, given by the system of interactions (the potential) which underlies every Gibbs distribution, turns the problem of identification to a standard parameter estimation problem. Parameter estimation for Gibbs distributions is usually based on the "maximum likelihood" (ML) approach, and its "maximum entropy" or "minimum distance" modifications (cf. e. g. Geman and Geman [6], Gidas [10], Younès [26], Janžura [17] for the general case, and Künsch [21], Janžura [18] for the special Gaussian case). The ML estimate is theoretically well understood, its consistency (see above and for the special Ising model case also Janžura [16]) can be proved in general, while inside the uniqueness region (with no phase transitions) also the asymptotic normality and efficiency is proven (cf. Gidas [10], Janžura [17]). Unfortunately, the numerical feasibility of the ML method is strongly limited, therefore its implementation is rather intricate, and only

[1]The research is supported by the Grant Agency of the Czech Republic under Grant No. 202/96/0731.

some approximations (cf. e.g. Strauss [25], Janžura [16]) or sophisticated stochastic approximation methods (Younès [26] and [27]) are available.

The implementation problem enforced a search for some modified methods which would keep, to some extend, the advantages of the ML method while avoiding its disadvantages. This effort was initiated with the "coding method" (Besag [1]), and finally led into the "maximum pseudo-likelihood" (MPL) method (Besag [2], Geman and Graffigne [7], Gidas [9], Guyon [12]) which consists in replacing the multidimensional joint distributions with the a product of conditional distributions which can be easily evaluated for the Gibbs distributions, providing we mean a conditional distribution of a rather small number (range) of variables with the all others being given. The consistency of the estimate remains true (cf. Geman and Graffigne [7], Gidas [9]).

In the present paper the major attention is paid to the problem of asymptotic normality and efficiency. After some preparatory sections with results adapted mostly from Georgii [8] and Preston [24], a whole class (with various ranges of the particular conditional distributions under consideration, although only rather small ranges are relevant for practical purposes) of MPL estimates together with the original ML estimate are defined in Section 6. We prove the consistency of every particular estimator (Theorem 6.1), and, moreover, we show that for every fixed sample size the ML estimate could be approximated by a sequence of MPL estimates. As it will be demonstrated, once the empirical distributions are introduced (Section 4) and consequently used, all the proofs become rather straightforward.

In the following Section 7 we proceed from the "optimization" version of the problem to the "normal equations" approach where the involvement of the phase transitions in the problem of investigating finer asymptotic properties is more transparent. Nevertheless, generalizing a central limit theorem obtained by Guyon and Künsch [22] for the particular case of the Ising model, we obtain the asymptotic normality of the MPL estimate, providing the true Gibbs distribution is ergodic, i.e. for the "pure phases". For a stationary Gibbs distribution, i.e. for a "mixture" of the pure phases, we obtain the limiting distribution as a corresponding mixed normal.

As the asymptotic efficiency is concerned, it is well known (cf. e.g. Hájek [14]) that a straightforward comparison of the asymptotic variances (cf. Guyon and Künsch [13] for the Ising model and Gidas [10] in general) is not satisfactory without some regularity conditions being satisfied. Thus, the problem of efficiency can be properly studied only inside the uniqueness region where "everything is smooth" and the central limit theorem holds in general for every finite range functional (Section 8). Moreover, the parameter family of uniquely defined Gibbs distribution obeys the regularity condition of local asymptotic normality (cf. Proposition 8.1 iii), and the asymptotic efficiency of the ML estimate follows in a rigorous way.

Since the asymptotic normality of the MPL estimates has been already proved (no mixtures here), we can directly observe a natural decrease of the asymptotic efficiency. However, for growing range of the MPL estimator the maximum asymptotic efficiency is approached (Theorem 9.5). The proof of this result in Section 10 involves some necessary techniques which are adopted from Künsch [22]. A short

remark on the infinitesimal robustness is also included (Section 11).

2. GIBBS RANDOM FIELDS

Considering a finite state space X with the σ-algebra of all its subsets $\mathcal{F} = \exp X$, by the random field (r.f.) we mean a probability measure μ on the product measurable space $(X, \mathcal{F})^T$ where the index set T is given by the d-dimensional ($d \geq 1$) discrete lattice Z^d.

We shall denote $x_A = \mathrm{Proj}_A(x)$ and $\mathcal{B}(A) = \mathrm{Proj}_A^{-1}(\mathcal{F}^A)$ for every $x \in X^T$, $A \subset T$, where $\mathrm{Proj}_A : X^T \to X^A$ is the corresponding projection function. Sometimes we shall not distinguish between x_A and $\mathrm{Proj}_A^{-1}(x_A)$.

A r.f. μ is said to be stationary if it is translation invariant, i.e. $\mu\tau_t^{-1} = \mu$ for every shift τ_t, $t \in T$, defined through $[\tau_t(x)]_s = x_{t+s}$ for every $x \in X^T$, $s \in T$. A stationary r.f. μ is called ergodic if its restriction to the σ-algebra $\mathcal{S} = \{F \in \mathcal{F}^T : \tau_t F = F \text{ for every } t \in T\}$ of invariant sets assumes only values zero or one, i.e. for every $F \in \mathcal{S}$ it holds: if $\mu(F) > 0$ then $\mu(F) = 1$.

Let us denote $k(T;t) = \{A \subset T;\ A \ni t,\ 0 < |A| < \infty\}$ for every $t \in T$, $k(T) = \bigcup_{t \in T} k(T;t)$ (by $|A|$ we mean the cardinality of $A \subset T$).

For every $W \subset T$ we denote by C_W the set of $\mathcal{B}(W)$-measurable bounded real-valued functions. Note that $C_W \subset C(X^T)$ for every $W \in k(T)$, where $C(X^T)$ is the set of real-valued functions which are bounded and continuous with respect to the usual product topology on X^T.

The potential is a family $U = \{U_A\}_{A \in k(T)}$ with $U_A \in C_A$ for every $A \in k(T)$, every particular map U_A being called the interaction corresponding to the set $A \in k(T)$.

We shall deal only with the potentials with are stationary:
$U_A(x_A) = U_{A-t}(\tau_t(x)_{A-t})$ for every $A \in k(T)$, $t \in T$, $x \in X^T$; and bounded:
$\|U\| = \sum_{A \in k(T;0)} |A| \cdot \|U_A\|_\infty < \infty$, where $\|U_A\|_\infty = \sup_{x_A \in X^A} |U_A(x_A)|$.

The set $\mathcal{U}$ of bounded stationary potentials is a Banach space with the norm $\|\cdot\|$.

A potential is of a finite range if the interactions vanish for large index sets, i.e. there exists $r \geq 0$ such that $U_A \equiv 0$ for every A : $\mathrm{diam}(A) > r$. Obviously, any finite range potential is bounded. We denote by $\mathcal{U}_r$ the set of potentials with a fixed range $r \geq 0$.

For a fixed potential $U \in \mathcal{U}$ the specification Π^U is the family of maps

$$\Pi^U = \left\{\Pi_A^U : X^A \times X^{T \setminus A} \to [0,1]\right\}_{A \in k(T)},$$

each Π_A^U being defined through

$$\Pi_A^U\left(x_A | x_{T \setminus A}\right) = \left[Z_A^U(x_{T \setminus A})\right]^{-1} \cdot \exp\left\{F_A^U(x_A | x_{T \setminus A})\right\}$$

where

$$F_A^U\left(x_A | x_{T \setminus A}\right) = \sum_{V \in k(T),\, V \cap A \neq \emptyset} U_V(x_V)$$

and

$$Z_A^U(x_{T \setminus A}) = \sum_{y_A \in X^A} \exp\left\{F_A^U(y_A | x_{T \setminus A})\right\}$$

is the appropriate normalizing constant.

A r.f. μ is called Gibbs with respect to a potential $U \in \mathcal{U}$, we write $\mu \in G(U)$, if its family of finite-dimensional conditional distributions is given by the specification Π^U, i.e. $\mu \in G(U)$ iff $\mu(x_A|\mathcal{B}(T \setminus A)) = \Pi_A^U(x_A|\cdot)$ a.s. holds for every $A \in k(T)$ and $x_A \in X^A$. (Here by $\mu(x_A|\mathcal{B}(T \setminus A))$ we mean the conditional probability of the "set" $x_A \in \mathcal{B}(A)$ under the σ-algebra $\mathcal{B}(T \setminus A)$.)

3. PARAMETER FAMILY

On the Banach space $(\mathcal{U}, \|\cdot\|)$ there is an equivalence relation generated by the specifications. For $U, \overline{U} \in \mathcal{U}$ we shall write $U \approx \overline{U}$ (saying the potentials are equivalent) if the corresponding specifications Π^U, $\Pi^{\overline{U}}$ are equal. Thanks to basic properties of conditional distributions we can observe that $U \approx \overline{U}$ iff for some $A \in k(T)$ there exists a function $\rho_A : X^{T \setminus A} \to \mathcal{R}$ satisfying

$$\rho_A(x_{T \setminus A}) = F_A^U(x_A|x_{T \setminus A}) - F_A^{\overline{U}}(x_A|x_{T \setminus A})$$

for every $x \in X^T$, i.e. $F_A^U - F_A^{\bar{U}} \in C_{T \setminus A}$.

Potentials $U^1, \ldots, U^N \in \mathcal{U}$ are said to be mutually non-equivalent if their linear combination can be equivalent to the zero potential $\mathbf{0} = \{\mathbf{0}_V \equiv 0\}_{V \in k(T)}$ only if it is the zero one, i.e.

$$\text{if} \quad \sum_{i=1}^{N} c_i U^i \approx \mathbf{0} \quad \text{then } c_1 = \ldots c_N = 0.$$

Hence

$$\sum c_i F_A^{U^i}(\cdot|\cdot) \in C_T \setminus C_{T \setminus A}$$

for every fixed $A \in k(T)$ and nonzero $(c_1, \ldots, c_N)^\top \in R^N$.

Let us mention that the mutual non-equivalence is the regularity (identifiability) condition here , which is a bit more complicated due to the infinite dimensional space X^T. This can be also seen from the preceding claim which is a generalization of the standard regularity condition requiring simply a linearly independent basis in exponential families.

For the sake of brevity we denote $F_A = (F_A^1, \ldots, F_A^N)^\top$ where we write F_A^i instead of $F_A^{U^i}$ for every $A \in k(T)$ and $i = 1, \ldots, N$. Let us realize that for every $U \in \mathcal{U}_r$ with fixed $r > 0$, we have $F_A^U(\cdot|\cdot) \in C_{\partial A}$, where

$$\partial A = A \cup \bigcup_{V \cap A \neq \emptyset,\, \mathrm{diam}(V) \leq r} V.$$

Similarly, we denote $\boldsymbol{g} = (g^1, \ldots, g^N)^\top$ where

$$g^j(x) = \sum_{V \in k(T;0)} |V|^{-1} U_V^j(x_V)$$

for every $x \in X^T$ and $j = 1, \ldots, N$.

Now, suppose we are given a fixed collection $U^1, \dots, U^N \in \mathcal{U}_r$ of mutually non-equivalent potentials of some fixed finite range $r > 0$.

Let us denote by $\mathcal{L} = \mathrm{Lin}\{U^1, \dots, U^N\}$ the finite dimensional subspace of $\mathcal{U}$ spanned by

$$U^1, \dots, U^N \in \mathcal{U}_r.$$

Thus, there is a one-to-one correspondence between the potential $U \in \mathcal{L}$ and the specification Π^U generated by U.

Moreover, there is the well-known isomorphism $\Phi : \mathcal{L} \to \mathcal{R}^N$ between the N-dimensional Banach space $\mathcal{L}$ and the N-dimensional Euclidean space, i.e.

$$\Phi(U) = \theta = (\theta_1, \dots, \theta_N)^\top \in \mathcal{R}^N \quad \text{iff} \quad U = \sum_{i=1}^{N} \theta_i U^i \in \mathcal{L}.$$

For every $\theta = \Phi(U) \in \mathcal{R}^N$ we shall write Π^θ and $G(\theta)$ instead of Π^U and $G(U)$, respectively, and we shall deal with the parameter family

$$\{G_I(\theta)\}_{\theta \in \mathcal{R}^N},$$

where $G_I(\theta)$ is the class of stationary Gibbs r.f.'s with respect to the potential $U = \Phi^{-1}(\theta)$.

Similarly, by $G_E(\theta)$ we denote the class of ergodic Gibbs r.f.'s. Let us recall that $G_E(\theta) = \mathrm{ex}\, G_I(\theta)$, i.e. ergodic r.f.'s are the extremal measures in $G_I(\theta)$, and by the ergodic decomposition theorem (cf. e.g. Theorem 14.10 in Georgii [8]), for every $\mu^0 \in G_I(\theta^0)$ we obtain

$$\mu^0(\Omega) = \int_{G_E(\theta^0)} \nu(\Omega)\, \mathrm{d}P(\nu)$$

for every $\Omega \in \mathcal{F}^{\mathrm{T}}$, where P is a uniquely defined probability measure on the set $G_E(\theta^0)$ of the ergodic Gibbs r.f.'s with an appropriate σ-algebra.

Remark 3.1. The problem of equivalence of the potentials can be easily avoided by considering only the so called "vacuum" potentials. A potential $U \in \mathcal{U}$ is a vacuum potential, we write $U \in \mathcal{U}_b$ if for every $A \in k(T)$ it holds $U_A(x_A) = 0$ whenever $x_t = b$ for some $t \in A$. Here $b \in X$ is a fixed stated called vacuum. Then it can be easily observed that $U \approx \mathbf{0}$ means $U = \mathbf{0}$, i.e. the equivalence relation turns to the identity. For details cf. e.g. Dobrushin and Nahapetian [5].

Remark 3.2. Note that there may exist non-stationary Gibbs r.f.'s with respect to a stationary potential $U \in \mathcal{U}$, i.e. $G(U) \setminus G_I(U) \neq \emptyset$. This phenomenon is called the breakdown of symmetry and makes the general study of Gibbs r.f.'s even more intricate (cf. also Remark 6.4 below).

4. EMPIRICAL RANDOM FIELDS

Suppose a collection of data $\hat{x}_{W_n} \in X^{W_n}$ obtained from an observation region $W_n \in k(T)$ to be generated by an unknown r.f. $\mu^0 \in G_I(\theta^0)$. We assume the observation

region W_n to be large enough to contain the lattice cube $V_n = [-n, n]^d \cap T$, i.e. $W_n \supset V_n$.

Thus, we may define the empirical r.f. $\mu^n_{\hat{x}}$ in a standard way, i.e.

$$\int f \, \mathrm{d}\mu^n_{\hat{x}} = |V_n|^{-1} \sum_{t \in V_n} f \circ \tau_t(\hat{x}^n_{\mathrm{per}})$$

for every bounded measurable f, where $\hat{x}^n_{\mathrm{per}}$ is the periodic continuation of $\hat{x}_{V_n}$. (We understand the empirical r.f. $\mu^n_{\hat{x}}$ to be defined for every $\hat{x} \in X^T$, being identical for all $\hat{y} \in \mathrm{Proj}^{-1}_{V_n}(\hat{x}_{V_n})$.) This is the "stationary version" of the empirical r.f. since really $\mu^n_{\hat{x}} \tau_t^{-1} = \mu^n_{\hat{x}}$ for every $t \in T$, and it is uniquely defined.

For deriving the second order properties we need the "unbiased version" $\tilde{\mu}^n_{\hat{x}}$ given by

$$\int f \, d\tilde{\mu}^n_{\hat{x}} = |V_n|^{-1} \sum_{t \in V_n} f \circ \tau_t(\hat{x})$$

for every $f \in C_V$ with $\bigcup_{t \in V_n}(t + V) \subset W_n$. We can see that the actual knowledge of $\hat{x}_{W_n}$ is sufficient in this case but the r.f. is not determined completely. On the other hand, the uniquely defined quantities $\int f \, \mathrm{d}\tilde{\mu}^n_{\hat{x}}$, $f \in C_V$, are usually sufficient for our purposes. Moreover, since really

$$\int \left[\int f \, d\tilde{\mu}^n_{\hat{x}} \right] \mathrm{d}\mu(\hat{x}) = \int f \, \mathrm{d}\mu$$

holds for every stationary r.f. μ, its name is fairly justified.

Further, we observe

$$\left| \int f \, \mathrm{d}\tilde{\mu}^n_{\hat{x}} - \int f \, \mathrm{d}\mu^n_{\hat{x}} \right| \leq \|f\|_\infty |V_n|^{-1} \, |\{t \in V_n; \, (t + V) \not\subset V_n\}| \longrightarrow 0 \quad \text{for } n \to \infty.$$

Therefore the two versions are asymptotically equivalent (uniformly for every $\hat{x} \in X^T$).

The following lemma could be strengthened, but this version is fully satisfactory for our purposes.

Lemma 4.1. For every collection $f_1, \ldots, f_N \in C_V$, $V \in k(T)$, it holds

$$\lim_{n \to \infty} \min_{\mu \in G_I(\theta^0)} \max_{j=1,\ldots,N} \left| \int f_j \, \mathrm{d}\mu^n_{\hat{x}} - \int f_j \, \mathrm{d}\mu \right| = 0 \quad \text{a.s. } [\mu^0] \quad \text{for every } \mu^0 \in G_I(\theta^0).$$

Proof. Let us denote

$$\Omega = \left\{ \hat{x} \in X^T; \limsup_{n \to \infty} \min_{\mu \in G_I(\theta^0)} \max_{j=1,\ldots,N} \left| \int f_j \, \mathrm{d}\mu^n_{\hat{x}} - \int f_j \, \mathrm{d}\mu \right| > 0 \right\} \in \mathcal{F}^{\mathrm{T}}.$$

Then $\nu(\Omega) = 0$ for every $\nu \in G_E(\theta^0)$ by the multidimensional ergodic theorem (cf. e.g. Theorem 14.A8 in Georgii [8]) which ensures for every $j = 1, \ldots, N$

$$\int f_j \, \mathrm{d}\mu^n_{\hat{x}} \longrightarrow \int f_j \, \mathrm{d}\nu \quad \text{a.s.}[\nu].$$

Thus $\mu^0(\Omega) = 0$ by the ergodic decomposition (cf. Section 3). □

Remark 4.2. It is worth mentioning that under the topology of local convergence (which here coincides with the usual weak topology – cf. e. g. Georgii [8] for details) $G_I(\theta^0)$ is a compact convex set and the map $\nu \mapsto \int f \, d\nu$ is continuous. Therefore the minima in the preceding lemma are actually attained, although they are random elements depending on the particular $\hat{x}$.

5. SOME THERMODYNAMICS

We denote by

$$p(U) = \lim_{n\to\infty} |V_n|^{-1} \log Z^U_{V_n}(x_{T\setminus V_n})$$

the pressure corresponding to the potential $U \in \mathcal{U}$. Note that the limit exists uniformly for every $x \in X^T$ (cf. e. g. Theorem 15.30 in Georgii [8]).

Direct calculations can show that $|V_n|^{-1} \log Z^U_{V_n}(x_{T\setminus V_n})$ is a convex function of $U \in \mathcal{U}$ and, moreover, it satisfies

$$|V_n|^{-1} \left| \log Z^{U^1}_{V_n}(x_{T\setminus V_n}) - \log Z^{U^2}_{V_n}(x_{T\setminus V_n}) \right| \le \|U^1 - U^2\|$$

for every U^1, $U^2 \in \mathcal{U}$, uniformly for every positive integer n and $x \in X^T$.

Therefore the same remains true also for the limiting function p.

In what follows, we shall deal with the restriction of p to the subspace $\mathcal{L}$. We sha" write $Z^\theta_{V_n}$ and $p(\theta)$ instead of $Z^U_{V_n}$ and $p(U)$, respectively, for $\theta = \Phi(U)$, $U \in \mathcal{L}$, and we shall understand $p/\mathcal{L}$ as a real-valued function on the space $\mathcal{R}^N$ equipped with the standard Euclidean norm $\|\cdot\|_2$.

Let us also recall that the strong convexity is a stronger convex property ensuring e. g. positive second derivative whenever it exists (see Section 8 below).

Lemma 5.1.

i) For every θ^1, $\theta^2 \in \mathcal{R}^N$ it holds

$$|V_n|^{-1} \left| \log Z^{\theta^1}_{V_n}(x_{T\setminus V_n}) - \log Z^{\theta^2}_{V_n}(x_{T\setminus V_n}) \right| \le \text{const} \cdot \|\theta^1 - \theta^2\|_2$$

uniformly for every positive integer n and $x \in X^T$ and

$$|p(\theta^1) - p(\theta^2)| \le \text{const} \cdot \|\theta^1 - \theta^2\|_2.$$

ii) The pressure $p : \mathcal{R}^N \to \mathcal{R}$ is a strictly convex continuous function. On every compact $K \subset \mathcal{R}^N$ it is even strongly convex.

Proof. The assertion i) follows from the above considerations. The properties of the pressure in ii) follow e. g. from Proposition 16.1 in Georgii [8] together with Dobrushin and Nahapetian [5]. □

Further important estimates are given in the following lemma.

Lemma 5.2. For a sequence $\{A_n\}_{n=1}^{\infty}$, $A_n \in k(T)$, with $\lim_{n\to\infty} |V_n|^{-1}|A_n \triangle V_n| = 0$ there exists a sequence of constants $c_n \to 0$ for $n \to \infty$ satisfying

i) $\left||A_n|^{-1} \log \Pi^{\theta}_{A_n}(x_{A_n}|y_{T\setminus A_n}) + p(\theta) - |V_n|^{-1} \sum_{t\in V_n} \theta^{\top} \boldsymbol{g} \circ \tau_t(x)\right| \le \|\theta\|_2\, c_n$ and

ii) $\left||A_n|^{-1} \log \mu(\hat{x}_{A_n}) + p(\theta) - \int \theta^{\top} \boldsymbol{g}\, \mathrm{d}\mu^n_{\hat{x}}\right| \le \|\theta\|_2\, c_n$

for every x, $\hat{x}$, $y \in X^T$, $\theta \in \mathcal{R}^N$, $\mu \in G_I(\theta)$.

Proof. The bounds can be deduced e. g. from results of Section 15.3 in Georgii [8] together with the properties of the empirical r.f.'s given in Section 4. □

The obtained results can be used to derive special forms of some important thermodynamic (or information theoretic – if prefered) characteristics, namely the entropy rate

$$H(\mu) = \lim_{n\to\infty} |V_n|^{-1} \int [-\log \mu(x_{V_n})]\, \mathrm{d}\mu(x)$$

existing for every stationary r.f. μ, and the relative entropy rate (asymptotic I-divergence, information gain) given for a pair ν, μ of stationary r.f.'s by

$$H(\nu|\mu) = \lim_{n\to\infty} |V_n|^{-1} \int \log \frac{\nu(x_{V_n})}{\mu(x_{V_n})}\, \mathrm{d}\nu(x)$$

whenever the expressions make sense and the limit exists.

Thus, for $\mu \in G_I(\theta)$ and a stationary ν it holds

$$H(\mu) = p(\theta) - \int \theta^{\top} \boldsymbol{g}\, \mathrm{d}\mu \ge 0,$$

and

$$H(\nu|\mu) = p(\theta) - \int \theta^{\top} \boldsymbol{g}\, \mathrm{d}\nu - H(\nu) \ge 0$$

with equality in the latter expression iff $\nu \in G_I(\theta)$, this result being called the variational principle.

For $\nu \in G_I(\theta^1)$ we obtain

$$H(\nu|\mu) = p(\theta) - p(\theta^1) - \int (\theta - \theta^1)^{\top} \boldsymbol{g}\, \mathrm{d}\nu$$

with $H(\nu|\mu) = 0$ iff $\theta = \theta^1$, since $G_I(\theta) \cap G_I(\theta^1) = \emptyset$ for $\theta \neq \theta^1$ e. g. by Theorem 2.34 in Georgii [8].

Consequently,

$$\min_{\theta\in\mathcal{R}^N} \left[p(\theta) - \int \theta^{\top} \boldsymbol{g}\, \mathrm{d}\nu\right]$$

is attained at the single point θ^1. (For more detailed treatment cf. e. g. Georgii [8].)

We finish this section with an important general lemma. We denote $\sigma_\delta(\theta^0) = \{\theta \in \mathcal{R}^N;\ \|\theta - \theta^0\|_2 \le \delta\}$ for every $\delta > 0$.

Lemma 5.3. Let $\{p_n\}_{n=1}^{\infty}$ be a sequence of real-valued continuous functions on $\mathcal{R}^N$. Let Q_{θ^0} be a class of functions satisfying

$$q(\theta) - q(\theta^0) \geq \gamma_\delta ||\theta - \theta^0||_2$$

with some $\gamma_\delta > 0$ for every $\delta > 0$, $\theta \notin \sigma_\delta(\theta^0)$, and $q \in Q_{\theta^0}$.

Let us suppose that for every $\varepsilon > 0$ and sufficiently large $n \geq n_\varepsilon$ there exists $q_n \in Q_{\theta^0}$ with $|p_n(\theta) - q_n(\theta)| \leq \varepsilon ||\theta||_2$ for every $\theta \in \mathcal{R}^N$.

Then for every $\delta > 0$ and sufficiently large $n \geq n_\delta$

$$\min_{\theta \in \mathcal{R}^N} p_n(\theta)$$

is attained at some $\theta^n \in \sigma_\delta(\theta^0)$, and therefore $\theta^n \to \theta^0$ for $n \to \infty$.

Proof. Let us choose $\delta > 0$ and set $\varepsilon = \frac{\gamma_\delta \cdot \delta}{2(\delta + ||\theta^0||_2)}$.

Then for $n \geq n_\varepsilon$ and $\theta \notin \sigma_\delta(\theta^0)$ we have

$$p_n(\theta) \geq q_n(\theta) - \varepsilon ||\theta||_2 \geq q_n(\theta^0) + \gamma_\delta ||\theta - \theta^0||_2 - \varepsilon ||\theta||_2 > q_n(\theta^0) + \varepsilon ||\theta^0||_2 \geq p_n(\theta^0).$$

Since p_n is continuous, its minimum must be attained at some θ^n inside $\sigma_\delta(\theta^0)$, and $\theta^n \to \theta^0$ is obvious. □

6 MAXIMUM LIKELIHOOD AND MAXIMUM PSEUDO-LIKELIHOOD ESTIMATION

The maximum likelihood estimate $\overline{\theta}^n$ of the parameter $\theta^0 \in \mathcal{R}^N$ based on the data collection $\hat{x}_{W_n} \in X^{W_n}$ should be in a rigid way defined by

$$\overline{\theta}^n = \operatorname*{argmax}_{\theta \in \mathcal{R}^N} \max_{\mu \in G_I(\theta)} \left\{ |W_n|^{-1} \log \mu(\hat{x}_{W_n}) \right\}.$$

Here the maximum over $G_I(\theta)$ is added in order to follow strictly the principle of seeking for the parameter corresponding to the most likely distribution.

However, Lemma 5.2 ii) provides us with a convenient approximation, and therefore we shall understand under the maximum likelihood estimate (MLE) its approximate version

$$\hat{\theta}^n = \operatorname*{argmax}_{\theta \in \mathcal{R}^N} \left\{ \int \theta^\top \boldsymbol{g} \, \mathrm{d}\mu_{\hat{x}}^n - p(\theta) \right\}.$$

This kind of estimate can be also derived from the "minimum distance principle", since it is obtained by minimizing the relative entropy

$$H(\mu_{\hat{x}}^n | \mu) = p(\theta) - \int \theta^\top \boldsymbol{g} \, \mathrm{d}\mu_{\hat{x}}^n - H(\mu_{\hat{x}}^n)$$

of the empirical r.f. $\mu_{\hat{x}}^n$ with respect to the theoretical $\mu \in G_I(\theta)$, where the entropy rate $H(\mu_{\hat{x}}^n)$ does not depend on the unknown parameter and therefore can be omitted.

Further, for every $A \in k(T)$ we define the corresponding maximum pseudo-likelihood estimate (MPLE) by

$$\hat{\theta}_A^n = \underset{\theta \in \mathcal{R}^N}{\operatorname{argmax}} \left\{ |A|^{-1} \int \log \Pi_A^\theta(x_A | x_{T \setminus A}) \, \mathrm{d}\mu_{\hat{x}}^n(x) \right\}.$$

This estimate may be also re-formulated in information theoretic terms, namely it minimizes the "mean relative conditional entropy"

$$\int H_0 \left([\mu_{\hat{x}}^n]_A(\cdot | x_{T \setminus A}) \middle| \Pi_A^\theta(\cdot | x_{T \setminus A}) \right) \mathrm{d}\mu_{\hat{x}}^n(x)$$

where $[\mu_{\hat{x}}^n]_A(\cdot|\cdot)$ is the corresponding conditional distribution derived from the empirical r.f. $\mu_{\hat{x}}^n$, and H_0 is the usual relative entropy. (Note that the empirical distributions and their entropies are well defined because the state space X is finite. For a general case we should proceed more carefully, but the idea would remain the same.)

Theorem 6.1.

i) The MLE $\hat{\theta}^n$ is defined with probability tending to one and it is consistent, i.e. for every $\mu^0 \in G_I(\theta^0)$

$$\mu^0 \left(\hat{x} \in X^T ; \max_{\theta \in \mathcal{R}^N} \left\{ \int \theta^\top \boldsymbol{g} \, \mathrm{d}\mu_{\hat{x}}^n - p(\theta) \right\} \text{ is attained} \right) \longrightarrow 1$$

and

$$\hat{\theta}^n \longrightarrow \theta^0 \quad \text{a.s. } [\mu^0] \text{ for } n \to \infty.$$

ii) For every $A \in k(T)$ the MPLE $\hat{\theta}_A^n$ is defined with probability tending to one and it is consistent, i.e. for every $\mu^0 \in G_I(\theta^0)$

$$\mu^0 \left(\hat{x} \in X^T ; \max_{\theta \in \mathcal{R}^N} \left\{ |A|^{-1} \int \log \Pi_A^\theta(\cdot|\cdot) \mathrm{d}\mu_{\hat{x}}^n \right\} \text{ is attained} \right) \longrightarrow 1$$

and

$$\hat{\theta}_A^n \longrightarrow \theta^0 \quad \text{a.s. } [\mu^0] \text{ for } n \to \infty.$$

Proof. For i) we set

$$Q_{\theta^0} = \left\{ q_\mu(\theta) = p(\theta) - \int \theta^\top \boldsymbol{g} \, \mathrm{d}\mu \right\}_{\mu \in G_I(\theta^0)}$$

and

$$p_n(\theta) = p(\theta) - \int \theta^\top \boldsymbol{g} \, \mathrm{d}\mu_{\hat{x}}^n \quad \text{for every } n \text{ and fixed } \hat{x} \in X^T.$$

For a.e. $x \in X^T[\mu^0]$ the assumptions of Lemma 5.3 are satisfied since by Lemma 5.2 ii) together with Remark 4.2 and the variational principle in Section 6 we obtain

$$\min_{\|\theta - \theta^0\|_2 = \delta} \min_{\mu \in G_I(\theta^0)} \left(q_\mu(\theta) - q_\mu(\theta^0) \right) = \delta \cdot \gamma_\delta > 0$$

and therefore by convexity $q_\mu(\theta)-q_\mu(\theta^0) \geq \gamma_\delta \|\theta-\theta^0\|_2$ for every $\theta \notin \sigma_\delta(\theta^0)$. Further from Lemma 4.1 we obtain

$$|p_n(\theta) - q_\mu(\theta)| < \|\theta\|_2 \cdot \varepsilon \quad \text{for some } \mu \in G_I(\theta^0) \text{ and large enough } n.$$

Moreover, the a.s. convergence in Lemma 4.1 yields the convergence in probability, and therefore the latter estimate, which guarantees the existence of minima, holds with probability tending to one.

For ii) we set

$$Q_{\theta^0} = \left\{ q_\mu(\theta) = -\int |A|^{-1} \log \Pi_A^\theta(\cdot|\cdot)\, \mathrm{d}\mu \right\}_{\mu \in G_I(\theta^0)}$$

and

$$p_n(\theta) = -\int |A|^{-1} \log \Pi_A^\theta(\cdot|\cdot)\, \mathrm{d}\mu_{\hat{x}}^n.$$

Since

$$q_\mu(\theta) - q_\mu(\theta^0) = |A|^{-1} \int \log \frac{\Pi_A^{\theta^0}(\cdot|\cdot)}{\Pi_A^\theta(\cdot|\cdot)}\, \Pi_A^{\theta^0}(\cdot|\cdot)\, \mathrm{d}\mu > 0$$

for every $\theta \neq \theta^0$ we obtain Q_{θ^0} to be a collection of strictly convex continuous functions with the minimum at θ^0. Thus the assumptions on Q_{θ^0} are satisfied.

Further it holds (note $\Pi_A^0(\cdot|\cdot) = \text{const.}$)

$$\begin{aligned} |q_\mu(\theta) - p_n(\theta)| &= \left| \int |A|^{-1} \left[\log \Pi_A^0(\cdot|\cdot) - \log \Pi_A^\theta(\cdot|\cdot) \right] (\mathrm{d}\mu - \mathrm{d}\mu_{\hat{x}}^n) \right| \\ &\leq \|\theta\|_2 \cdot \text{const} \cdot \max_{x_{\partial A} \in X^{\partial A}} |\mu(x_{\partial A}) - \mu_{\hat{x}}^n(x_{\partial A})| \leq \varepsilon \cdot \|\theta\|_2 \end{aligned}$$

for large enough n and some $\mu \in G_I(\theta^0)$ by Lemma 5.1 i), Lemma 4.1, and obvious uniform bound $\||A|^{-1} F_A^U(\cdot|\cdot)\| \leq \|U\|$. Thus the assumptions of Lemma 5.3 are satisfied, and the proof is completed in the same way as for i). □

Now let us fix the empirical r.f. $\mu_{\hat{x}}^n$. We can study the behaviour of the MPLE $\hat{\theta}_A^n$ for growing A. For this purpose let $\{A_k\}_{k=1}^\infty$ satisfy the assumption of Lemma 5.2, namely let $|V_k|^{-1} |A_k \triangle V_k| \longrightarrow 0$ for $k \to \infty$.

Proposition 6.2. Let the MLE $\hat{\theta}^n$ exist. Then the MPLE $\hat{\theta}_{A_k}^n$ exists for sufficiently large k, and

$$\hat{\theta}_{A_k}^n \longrightarrow \hat{\theta}^n \quad \text{for } k \to \infty.$$

Proof. We set

$$Q_{\hat{\theta}^n} = \left\{ q(\theta) = p(\theta) - \int \theta^\top \boldsymbol{g}\, \mathrm{d}\mu_{\hat{x}}^n \right\}$$

which is now a singleton since the empirical distribution $\mu_{\hat{x}}^n$ is uniquely defined, and

$$p_k(\theta) = -\int |A_k|^{-1} \log \Pi_{A_k}^\theta(\cdot|\cdot)\, \mathrm{d}\mu_{\hat{x}}^n \quad \text{for every } k.$$

Then the statement follows from Lemma 5.3 and Lemma 5.2 i). □

Remark 6.3. From Lemma 5.3 ii) by Lemma 5.3 it even follows that the "true" MLE $\overline{\theta}^n$ is also defined (in sense of existence of local maxima for sufficiently large n) and consistent. We must only prove in addition that the multifunction

$$\theta \mapsto \{\mu(\hat{x}_{W_n})\}_{\mu \in G_I(\theta)}$$

really attains its maximum in every ball $\sigma_\delta(\theta^0)$. But, by the "compactness" arguments we obtain a sequence $\mu_j \in G_I(\theta^j)$ with $\theta^j \longrightarrow \theta^* \in \sigma_\delta(\theta^0)$ and $\mu_j \Longrightarrow \mu^*$ weakly for $j \to \infty$ for some stationary r.f. μ^*, satisfying

$$\mu_j(\hat{x}_{W_n}) \longrightarrow \sup_{\theta \in \sigma_\delta(\theta^0)} \max_{\mu \in G_I(\theta)} \mu(\hat{x}_{W_n}) = \mu^*(\hat{x}_{W_n}).$$

Since the entropy rate H is upper semicontinuous (cf. e.g. Proposition 15.14 in Georgii [8]) we obtain from the variational principle $\mu^* \in G_I(\theta^*)$.

Remark 6.4. The weak consistency of the estimates (i.e. the convergence in probability) can be proved with the aid of the appropriate large deviations theorems for the non-stationary Gibbs r.f.'s as well (cf. Gidas [10] and Comets [4]). An exponential rate of convergence consequently follows.

7. ASYMPTOTIC NORMALITY OF THE MAXIMUM PSEUDO-LIKELIHOOD ESTIMATES

Since

$$-\log \Pi_A^\theta(x_A|x_{T\setminus A}) = \log Z_A^\theta(x_{T\setminus A}) - \theta^\top F_A(x_A|x_{T\setminus A})$$

is a smooth convex function of $\theta \in \mathcal{R}^N$ we obtain an equivalent definition of the MPLE $\hat{\theta}_A^n$, namely

$$\hat{\theta}_A^n = \operatorname*{argmin}_{\theta \in \mathcal{R}^N} \left\{ -\int |A|^{-1} \log \Pi_A^\theta(\cdot|\cdot)\, \mathrm{d}\mu_{\hat{x}}^n \right\}$$

iff

$$J_{\mu_{\hat{x}}^n}^A(\hat{\theta}_A^n) = 0,$$

where for every stationary r.f. μ we define

$$J_\mu^A(\theta) = \int S_A^\theta \, \mathrm{d}\mu \quad \text{for every } \theta \in \mathcal{R}^N,$$

and

$$S_A^\theta = \left[\frac{\mathrm{d}}{\mathrm{d}\theta_j} \{-|A|^{-1} \log \Pi_A^\theta(\cdot|\cdot)\} \right]_{j=1,\ldots,N} = |A|^{-1} \left[E_A^\theta[F_A](\cdot) - F_A(\cdot|\cdot) \right]$$

with

$$E_A^\theta[F_A^i](x_{T\setminus A}) = \sum_{y_A \in X^A} F_A^i(y_A|x_{T\setminus A})\, \Pi_A^\theta(y_A|x_{T\setminus A})$$

for every $i = 1, \ldots, N$ and $x_{T\setminus A} \in X^{T\setminus A}$.

Further we define

$$\begin{aligned} D_A^\theta(x) &= \nabla S_A^\theta(x) = \left(\frac{\mathrm{d}}{\mathrm{d}\theta_j}[S_A^\theta(x)]^i\right)_{i,j=1,\dots,N} \\ &= |A|^{-1}\left[\mathrm{cov}_A^\theta(F_A^i, F_A^j)(x_{T\setminus A})\right]_{i,j=1,\dots,N} \geq 0 \end{aligned}$$

for every $\theta \in \mathcal{R}^N$ and $x \in X^T$, where

$$\begin{aligned} \mathrm{cov}_A^\theta(F_A^i, F_A^j)(x_{T\setminus A}) = &\sum_{y_A \in X^A} F_A^i(y_A|x_{T\setminus A})\, F_A^j(y_A|x_{T\setminus A})\, \Pi_A^\theta(y_A|x_{T\setminus A}) \\ &- E_A^\theta[F_A^i](x_{T\setminus A}) \cdot E_A^\theta[F_A^j](x_{T\setminus A}). \end{aligned}$$

Lemma 7.1. Let μ be a positive stationary r.f. (i.e. $\mu(x_B) > 0$ for every $x_B \in \mathcal{B}(B)$, $B \in k(T)$). Then J_μ^A is a one-to-one regular mapping with positive definite Jacobi matrix

$$\nabla J_\mu^A(\theta) = \left(\frac{\mathrm{d}}{\mathrm{d}\theta_j}[J_\mu^A(\theta)]^i\right)_{i,j=1,\dots,N} > 0$$

at every $\theta \in \mathcal{R}^N$.

Proof. By definition it holds $\nabla J_\mu^A(\theta) = \int D_A^\theta \,\mathrm{d}\mu$ where the matrix $D_A^\theta(x)$ is in general positive semidefinite for every $A \in k(T)$, $\theta \in \mathcal{R}^N$, and $x \in X^T$. Since $U^1, \dots, U^N$ are mutually non-equivalent, for every $0 \neq c = (c_1, \dots, c_N)^\top \in R^N$ there exists $x \in X^T$ with $c^\top D_A^\theta(x)\, c > 0$ (cf. Section 3). But $c^\top D_A^\theta(x)\, c \in C_{\partial A}$ and μ is positive, therefore $\int D_A^\theta \mathrm{d}\mu > 0$. □

Remark 7.2. Accordingly, the inverse mapping $[J_\mu^A]^{-1}$ exists with similar properties. (Note that every $\mu \in \bigcup_{\theta \in R^N} G_I(\theta)$ and $\mu_{\hat{x}}^n$ for large enough n are positive.) Therefore the MPLE $\hat{\theta}_A^n$ can be defined by

$$\hat{\theta}_A^n = \left[J_{\mu_{\hat{x}}^n}^A\right]^{-1}(0),$$

whenever $J_{\mu_{\hat{x}}^n}^A$ is regular and 0 is contained in the open set $J_{\mu_{\hat{x}}^n}^A(\mathcal{R}^N)$. Since $S_A^\theta \in C_{\partial A}$ are uniformly bounded for every $\theta \in \mathcal{R}^N$ we obtain $J_{\mu_j}^A \rightrightarrows J_\mu^A$ uniformly if $\mu_j \Longrightarrow \mu$ weakly. Moreover, if every $J_{\mu_j}^A$ is regular, the pointwise convergence of the inverse transforms can be concluded. We could follow this approach to prove the existence and the consistency of the MPLE $\hat{\theta}_A^n$.

Unfortunately, for the MLE estimate $\hat{\theta}^n$ we must proceed more carefully since the problem of phase transitions ($\theta \in \mathcal{R}^N$ with $|G_I(\theta)| > 1$) can not be avoided. Namely, it holds by the variational principle that the MLE

$$\hat{\theta}^n = \operatorname*{argmax}_{\theta \in \mathcal{R}^N} \left[\int \theta^\top \boldsymbol{g}\, \mathrm{d}\mu_{\hat{x}}^n - p(\theta)\right]$$

is given iff there exists some $\mu \in G_I(\hat{\theta}^n)$ satisfying

$$J^{\infty}_{\mu^n_{\hat{x}}}(\mu) = \int \left[\boldsymbol{g} - \int \boldsymbol{g}\, \mathrm{d}\mu \right] \mathrm{d}\mu^n_{\hat{x}} = 0.$$

Hence the estimate exists iff

$$\int \boldsymbol{g}\, \mathrm{d}\mu^n_{\hat{x}} \in \left\{ \int \boldsymbol{g}\, \mathrm{d}\mu;\ \mu \in G_I(\theta),\ \theta \in \mathcal{R}^N \right\} = \mathcal{G}.$$

The proof of Theorem 6.1 shows that $\mathcal{G}$ is an open subset of $\mathcal{R}^N$. According to Proposition 5.18 and Example 5.20 (1) in Georgii [8] the limit exists

$$\lim_{k \to \infty} J^{A_k}_{\mu}(\theta) = J^{\infty}_{\mu}(\nu_\theta(\mu)) \quad \text{for every } \theta \in \mathcal{R}^N,$$

where the particular $\nu_\theta(\mu) \in G_I(\theta)$ depends on the actual fixed stationary r.f. μ.

Obviously, $J^{\infty}_{\mu^n_{\hat{x}}}(\nu_\theta(\mu^n_{\hat{x}}))$ could be understood as a well-defined function of $\theta \in \mathcal{R}^N$, but it may not be continuous at the points of phase transitions, etc., and therefore this way seems useless. On the other hand, we can directly introduce the inverse transform

$$J^{-1}_{\mu^n_{\hat{x}}}(\lambda) = \left\{ \theta \in \mathcal{R}^N;\ \min_{\mu \in G_I(\theta)} \left\| \int \boldsymbol{g}\, \mathrm{d}\mu - \int \boldsymbol{g}\, \mathrm{d}\mu^n_{\hat{x}} - \lambda \right\|_2 = 0 \right\},$$

which is a well-defined continuous mapping, and finally we obtain $J^{-1}_{\mu_j} \longrightarrow J^{-1}_{\mu}$ for $\mu_j \Longrightarrow \mu$, and $[J^{A_k}_{\mu}]^{-1} \longrightarrow J^{-1}_{\mu}$ for $k \to \infty$.

And this is in fact the definite essence of Theorem 6.1 and Proposition 6.2.

The main aim of the present section consists in proving the asymptotic (mixed-) normality of the MPLE. Therefore we need an appropriate version of the central limit theorem.

For $f^1,\ f^2 \in C(X^T)$ and stationary r.f. μ we denote

$$B_\mu(f^1, f^2) = \sum_{t \in T} \mathrm{cov}_\mu(f^1, f^2 \circ \tau_t)$$

whenever the sum converges.

For $\boldsymbol{f} = (f^1, \ldots, f^N)^\top$ and $\boldsymbol{h} = (h^1, \ldots, h^N)^\top$ we denote

$$B_\mu[\boldsymbol{f}; \boldsymbol{h}] = (B_\mu(f^i, h^j))_{i,j=1,\ldots,N}.$$

For $T^1,\ T^2 \subset T$ we denote $T^1 \ominus T^2 = \{t - s;\ t \in T^1,\ s \in T^2\}$.

Lemma 7.3. Let $\mu \in G_I(\theta)$. Then

i) $B_\mu[S^\theta_A;\ S^\theta_A] = \sum_{t \in \partial A \ominus \partial A} \mathrm{cov}_\mu[S^\theta_A;\ S^\theta_A \circ \tau_t] > 0,$

and

ii) $-B_\mu[S^\theta_A; \boldsymbol{g}] = \int D^\theta_A \mathrm{d}\mu = \nabla J^A_\mu(\theta) > 0.$

Proof. For a fixed $c \neq 0$ we set $Y = c^\top S_A^\theta \in C_{\partial A}$, and define the potential $U^Y = \{U_B^Y\}_{B \in k(T)}$, where $U_B^Y = Y \circ \tau_t$ for $B = \partial A + t$, $t \in T$, and $U_B^Y \equiv 0$ otherwise. Let $U^Y \approx \mathbf{0}$. Then by the variational principle in Section 5 we obtain $\int Y \, d\nu = 0$ for every stationary r.f. ν, and especially for some $\nu^1 \in G_I(\theta^1)$, $\theta^1 = \theta + c$. Thus $0 = \int c^\top (S_A^{\theta^1} - S_A^\theta) \, d\nu^1 = c^\top \left[\int D_A^{\theta^*} \, d\nu^1\right] c$ where $\theta^* = \theta + \gamma c$, $\gamma \in [0,1]$, which contradicts Lemma 7.1. Therefore $U^Y \not\approx \mathbf{0}$.

Thus by Lemma 2 in Dobrushin and Nahapetian [5] we obtain a constant $\lambda > 0$ such that

$$|V_m|^{-1} \mathrm{cov}_{V_m}^\theta \left(F_{V_m}^{U^Y}, F_{V_m}^{U^Y}\right)(x_{T \setminus V_m}) \geq \lambda$$

for every $x \in X^T$ and some sequence of cubes $\{V_m\}_{m=1}^\infty$ with $V_m \nearrow T$ as $m \to \infty$.

Finally, utilizing the standard inequality $E[\mathrm{var}(\xi|\eta)] \leq \mathrm{var}(\xi)$, we obtain

$$\begin{aligned} 0 < \lambda &\leq |V_m|^{-1} E_\mu \left[\mathrm{cov}_{V_m}^\theta (F_{V_m}^{U^Y}, F_{V_m}^{U^Y})(\cdot)\right] \\ &\leq |V_m|^{-1} \mathrm{cov}_\mu \left(F_{V_m}^{U^Y}, F_{V_m}^{U^Y}\right) \\ &= \sum_{s \in \partial A \ominus \partial A} \mathrm{cov}^\mu (Y, Y \circ \tau_s) \cdot |(I_A^m + s) \cap I_A^m| \cdot |V_m|^{-1} \\ &\longrightarrow c^\top B_\mu [S_A^\theta; S_A^\theta] c \quad \text{as } m \to \infty, \end{aligned}$$

since $|V_m|^{-1} |(I_A^m + s) \cap I_A^m| \longrightarrow 1$ as $m \to \infty$ where $I_A^m = \{t \in T;\, (\partial A + t) \cap V_m \neq \emptyset\}$ for every m, and

$$\mathrm{cov}_\mu (Y, Y \circ \tau_k) = 0 \quad \text{if} \quad \partial A \cap (\partial A + t) = \emptyset.$$

Thus the statement i) is proved. The proof of ii) is a direct computation based on the observation that $\mathrm{cov}_\mu (S_A^\theta, f) = 0$ whenever $f \in C_{T \setminus A}$. □

Theorem 7.4. (CLT) Let $\mu \in G_E(\theta)$. Then

$$|V_n|^{-\frac{1}{2}} \sum_{t \in V_n} S_A^\theta \circ \tau_t \Longrightarrow \mathcal{N}_N(0, B_\mu [S_A^\theta; S_A^\theta]) \quad \text{in distribution } [\mu] \text{ as } n \to \infty.$$

Proof. Following Guyon and Künsch [13], for fixed $c \neq 0$ and every $t \in T$ we denote $Y_t = c^\top S_A^\theta \circ \tau_t$. We set $\xi_n = a_n^{-1} \sum_{t \in V_n} Y_t$ and $\xi_n^s = a_n^{-1} \sum_{t \in (\tilde{A}+s) \cap V_n} Y_t$ for every $s \in V_n$, where $\tilde{A} = \partial A \ominus \partial A$ and $a_n^2 = E_\mu \left[\sum_{t \in V_n} Y_t\right]^2$.

Then, due to e. g. Lemma 2 in Bolthausen [3], it is sufficient to verify

$$\lim_{n \to \infty} E_\mu \left[(i\lambda - \xi_n) e^{i\lambda \xi_n}\right] = 0 \quad \text{for every real } \lambda.$$

Employing the decomposition (again by Bolthausen [3])

$$(i\lambda - \xi_n) e^{i\lambda \xi_n} = C_{n,1} + C_{n,2} + C_{n,3}$$

where

$$C_{n,1} = -a_n^{-1} \sum_{t \in V_n} Y_t \, e^{i\lambda(\xi_n - \xi_n^t)},$$

$$C_{n,2} = -i\lambda \, e^{i\lambda \xi_n} \left(a_n^{-1} \sum_{t \in V_n} Y_t \, \xi_n^t - 1 \right),$$

and

$$C_{n,3} = e^{i\lambda \xi_n} a_n^{-1} \sum_{t \in V_n} Y_t \left(e^{-i\lambda \xi_n^t} + i\lambda \, \xi_n^t - 1 \right),$$

we observe $E_\mu \, C_{n,1} = 0$ for every n, $E_\mu |C_{n,2}| \longrightarrow 0$ as $n \to \infty$ by the mean ergodic theorem (cf. e. g. Theorem 14.A5 in Georgii [8] – here the assumption $\mu \in G_E(\theta)$ is needed), and $E_\mu |C_{n,3}| \longrightarrow 0$ as $n \to \infty$ by standard estimates. For details cf. Guyon and Künsch [13] for the particular Ising model or Janžura and Lachout [20] for the general case. □

Now, we can prove the asymptotic normality of the MPLE.

We shall consider the "unbiased version" $\tilde{\mu}_{\tilde{x}}^n$ of the empirical r. f. Thanks to the basic estimate

$$\left| \int f \, \mathrm{d}\mu_{\tilde{x}}^n - \int f \, d\tilde{\mu}_{\tilde{x}}^n \right| \leq \|f\|_\infty \, |V_n|^{-1} \, |\{t \in V_n;\, t + V \subset V_n\}| \longrightarrow 0 \quad \text{for } n \to \infty$$

for every $f \in C_V$ with $\bigcup_{t \in V_n} (t + V) \subset W_n$, the modification does not influence the problem of consistency of both the MLE and the MPLE. We shall quote the estimates based on the "unbiased version" $\tilde{\mu}_{\tilde{x}}^n$ of the empirical r. f. as the modified MLE and the modified MPLE, respectively.

Denoting the modified versions by $\tilde{\theta}^n$ and $\tilde{\theta}_A^n$ for $A \in k(T)$, respectively, we could replicate the proof of Theorem 6.1 to obtain

$$\tilde{\theta}^n \longrightarrow \theta^0 \quad \text{and} \quad \tilde{\theta}_A^n \longrightarrow \theta^0 \quad \text{a. s. } [\mu^{\theta^0}] \text{ for } n \to \infty \text{ and every } \theta^0 \in \Theta.$$

We must only keep in mind the assumed relation between V_n and W_n so that the empirical r. f. $\tilde{\mu}_{\tilde{x}}^n$ is defined for the particular $\boldsymbol{g}$ or S_A^θ, respectively. Such problem does not occur with the "stationary version" $\mu_{\tilde{x}}^n$ but its bias could break validity of the central limit theorems.

For the sake of brevity we denote

$$B_\theta^A(\mu) = B_\mu[\boldsymbol{g};\, S_A^\theta] \left(B_\mu[S_A^\theta;\, S_A^\theta] \right)^{-1} B_\mu[S_A^\theta;\, \boldsymbol{g}]$$

for $\mu \in G_I(\theta)$.

Theorem 7.5. For every $A \in k(T)$ the modified MPLE $\tilde{\theta}_A^n$ is asymptotically normal, providing the generating Gibbs r. f. is ergodic, namely for every $\mu \in G_E(\theta^0)$, $\theta^0 \in R^N$, it holds

$$|V_n|^{\frac{1}{2}} (\tilde{\theta}_A^n - \theta^0) \Longrightarrow \mathcal{N}_N \left(0,\, (B_\theta^A(\mu))^{-1}\right) \quad \text{in distribution } [\mu] \text{ as } n \to \infty.$$

Proof. By definition it holds

$$0 = \int S_A^{\tilde{\theta}_A^n} \mathrm{d}\tilde{\mu}_{\hat{x}}^n = \int S_A^{\theta^0} \mathrm{d}\tilde{\mu}_{\hat{x}}^n + \left[\int D_A^{\gamma^n\theta^0+(1-\gamma^n)\tilde{\theta}_A^n} \mathrm{d}\tilde{\mu}_{\hat{x}}^n\right](\tilde{\theta}_A^n - \theta^0)$$

with some $\gamma^n \in [0,1]$.

We observe

$$|V_n|^{\frac{1}{2}} \int S_A^{\theta^0} \,\mathrm{d}\tilde{\mu}_{\hat{x}}^n \Longrightarrow \mathcal{N}_N\left(0,\, B_\mu[S_A^{\theta^0}; S_A^{\theta^0}]\right)$$

by Theorem 7.4 since $\int \left[\int S_A^{\theta^0} \,\mathrm{d}\tilde{\mu}_{\hat{x}}^n\right] \mathrm{d}\mu(\hat{x}) = \int S_A^{\theta^0} \,\mathrm{d}\mu = 0$ thanks to the "unbiased version" of the empirical r.f. $\tilde{\mu}_{\hat{x}}^n$.

Further, since D_A^θ is uniformly bounded, $\tilde{\theta}_A^n$ is consistent estimate, and μ is an ergodic r.f., we obtain

$$\int D_A^{\gamma^n\theta^0(1-\gamma^n)\tilde{\theta}_A^n} \,\mathrm{d}\tilde{\mu}_{\hat{x}}^n \longrightarrow \int D_A^{\theta^0} \,\mathrm{d}\mu \quad \text{a.s. } [\mu].$$

The rest of the proof is standard. □

Corollary 7.6. For a stationary generating $\mu \in G_I(\theta^0)$ the modified MPLE $\tilde{\theta}_A^n$ is asymptotically mixed-normal, i.e.

$$|V_n|^{\cdot}(\tilde{\theta}_A^n - \theta^0) \Longrightarrow \int_{G_E(\theta^0)} \mathcal{N}_N\left(0,\, (B_\theta^A(\nu))^{-1}\right) \mathrm{d}P_\mu(\nu) \quad \text{in distribution } [\mu] \text{ as } n \to \infty,$$

where P_μ is the ergodic decomposition measure.

Proof. The result follows directly from the preceding Theorem 7.5 and the ergodic decompo ition. Namely, denoting $\eta_n = c^\top \left[|V_n|^{\frac{1}{2}}(\tilde{\theta}_A^n - \theta^0)\right]$, we obtain $\mu(\eta_n < \alpha) = \int_{G_E(\theta^0)} \nu(\eta_n < \alpha)\,\mathrm{d}P_\mu(\nu)$ for every n. By taking limit we obtain the claimed statement. □

8. DOBRUSHIN'S UNIQUENESS REGION

For every $t \in T$ let us define

$$\gamma_t(U) = \frac{1}{2} \sup \left\{ \sum_{x_0 \in X} \left| \Pi_{\{0\}}^U(x_0|y_{T\setminus\{0\}}) - \Pi_{\{0\}}^U(x_0|z_{T\setminus\{0\}}) \right| ;\, y_s = z_s \text{ for } s \neq t \right\}.$$

If $\gamma(U) = \sum_{t\in T} \gamma_t(U) < 1$ the potential is said to satisfy Dobrushin's condition. Gross [11] proved that the Dobrushin's uniqueness region $\mathcal{D} = \{U \in \mathcal{U};\, \gamma(U) < 1\}$ is an open subset of the space $\mathcal{U}$.

Moreover, for every $U^0 \in \mathcal{D}$ there exists an open neighborhood ∂U^0 satisfying

$$\gamma(\partial U^0) = \sum_{t\in T} \gamma_t(\partial U^0) < 1,$$

where

$$\gamma_t(\partial U^0) = \sup_{U \in \partial U^0} \gamma_t(U) \quad \text{for every } t \in T.$$

This is always possible, see the proof of Proposition 2 in Gross [11].

For every $U \in \mathcal{D}$ there is exactly one Gibbs r.f. μ^U (this is the famous Dobrushin's result – cf. e.g. Künsch [22], Corollary 2.3) which is, moreover, stationary and ergodic (cf. Theorem 4.1 and Theorem 4.3 in Preston [24]), i.e. $G(U) = G_I(U) = G_E(U) = \{\mu^U\}$ for every $U \in \mathcal{D}$.

We denote $\mathcal{E} = \mathcal{D} \cap \mathcal{L}$, where again $\mathcal{L} = \text{Lin}(U^1, \ldots, U^N)$ with mutually non-equivalent $U^1, \ldots, U^N \in \mathcal{U}_r$, $r > 0$.

For every $\theta \in \Phi(\mathcal{E}) = \Theta$ we shall write μ^θ instead of μ^U, and we shall deal with the parameter family

$$\mathcal{M} = \{\mu^\theta\}_{\theta \in \Theta}$$

of Gibbs r.f.'s with the open set of parameters $\Theta \subset \mathcal{R}^N$.

Let us note that Θ always contains the zero vector $\mathbf{0} \in R^N$, and $\mu^{\mathbf{0}}$ is in fact simply the corresponding infinite power of the uniform distribution.

Further, we may write directly

$$B_\theta(f^1, f^2) = \sum_{t \in T} \text{cov}^\theta(f^1, f^2 \circ \tau_t)$$

for every $f^1, f^2 \in C_W$ with $W \in k(T)$, and $\theta \in \Theta$, where cov^θ stands for the covariance with respect to μ^θ. The sum is now absolutely convergent by Theorem 5.1 in Künsch [22].

Moreover, again by Theorem 5.1 in Künsch [22] it holds

$$\frac{\mathrm{d}}{\mathrm{d}\theta_i} p(\theta) = \int g^i \, \mathrm{d}\mu^\theta$$

and

$$\frac{\mathrm{d}}{\mathrm{d}\theta_i} \int f \, \mathrm{d}\mu^\theta = B_\theta(f, g^i)$$

for every $i = 1, \ldots, N$; $f \in C_W$ with $W \in k(T)$, and $\theta \in \Theta$.

In particular we have

$$B_\theta[\boldsymbol{g}; \boldsymbol{g}] = \left(B_\theta(g^i, g^j)\right)_{i,j=1,\ldots,N} = \left(\frac{\mathrm{d}^2}{\mathrm{d}\theta_i \, \mathrm{d}\theta_j} p(\theta)\right)_{i,j=1,\ldots,N},$$

and we observe $B_\theta[\boldsymbol{g}; \boldsymbol{g}] > 0$, i.e. positive definite – by Lemma 5.1 ii) for every $\theta \in \Theta$.

Finally, let us denote by

$$\ell_n^\theta(x) = \left(\ell_n^\theta(x)^i = \frac{\mathrm{d}}{\mathrm{d}\theta_i} \log \mu^\theta(x_{V_n})\right)_{i=1,\ldots,N}$$

the corresponding score function.

Proposition 8.1.

i) For every $f \in C_W$, $W \in k(T)$, and $\theta \in \Theta$ it holds

$$|V_n|^{-\frac{1}{2}} \sum_{t \in V_n} \left[f \circ \tau_t - \int f \, \mathrm{d}\mu^\theta \right] \Longrightarrow \mathcal{N}_N(0, B_\theta(f,f))$$

for $n \to \infty$ in distribution $[\mu^\theta]$.

ii) For every $\theta \in \Theta$ is holds

$$|V_n|^{-\frac{1}{2}} \left\{ \sum_{t \in V_n} \left[g \circ \tau_t - \int g \, \mathrm{d}\mu^\theta \right] - \ell_n^\theta \right\} \longrightarrow 0 \quad \text{for } n \to \infty \text{ in probability } [\mu^\theta].$$

iii) The parameter family $\mathcal{M}$ obeys the regularity condition of local asymptotic normality (LAN), namely it holds

$$\begin{aligned} &\log \frac{\mu^{\theta + |V_n|^{-\frac{1}{2}}\alpha}(x_{V_n})}{\mu^\theta(x_{V_n})} \\ = \; &|V_n|^{-\frac{1}{2}} \sum_{t \in V_n} \alpha^{\mathrm{T}} \left[g \circ \tau_t(x) - \int g \, \mathrm{d}\mu^\theta \right] - \frac{1}{2} \alpha^{\mathrm{T}} B_\theta[g;g] \, \alpha + M_n^\theta(x) \end{aligned}$$

for every θ, $\theta + |V_n|^{-\frac{1}{2}}\alpha \in \Theta$, $x \in X^T$ and large enough positive integer n, where

$$M_n^\theta \longrightarrow 0 \quad \text{for } n \to \infty \text{ in probability } [\mu^\theta].$$

Proof. The first statement is the central limit theorem for functionals of Gibbs r.f.'s (cf. e.g. Theorem 4.1 in Künsch [22]).

The remaining statements can be obtained by appropriate expansions together with the bounds following from Theorem 3.2 in Künsch [22]. For details cf. Janžura [17] and [19]. □

Remark 8.2. The crucial central limit theorem in Proposition 8.1 i) can be also proved with replacing the assumption $\theta \in \Theta$ by a rather technical condition (cf. Theorem 2 in Gidas [10]) that guarantees the convergence of $B_\theta(f,f)$. Since for $\theta \in \Theta$ this condition is satisfied, the result seems to be more general. On the other hand it is not completely clear where else the condition can be satisfied in addition, and therefore we rather prefer the "uniqueness region" approach which will also provide us with some usefull bounds (cf. Section 10 below).

Moreover, the problem of positive definiteness of the asymptotic variance matrix, which is closely related to the strong convexity of the pressure (proved by Dobrushin and Nahapetian [5]), is not discussed in Gidas [10].

9. ESTIMATION IN THE UNIQUENESS REGION

We shall restrict our further considerations to the uniqueness region, namely to the parameter family

$$\mathcal{M} = \{\mu^\theta\}_{\theta\in\Theta}$$

of Gibbs r.f.'s. Again as in Section 7 we shall deal with the "unbiased" empirical r.f.'s and modified versions of the estimates.

Theorem 9.1. The modified MLE $\tilde{\theta}^n$ is asymptotically normal and asymptotically efficient, namely for every $\theta^0 \in \Theta$ it holds

$$|V_n|^{\frac{1}{2}}(\tilde{\theta}^n - \theta^0) \Longrightarrow \mathcal{N}_N\left(0, (B_{\theta^0}[\boldsymbol{g};\boldsymbol{g}])^{-1}\right) \quad \text{for } n \to \infty \text{ in distribution } [\mu^{\theta^0}]$$

and

$$|V_n|^{\frac{1}{2}}\left[\tilde{\theta}^n - \theta^0 - (B_{\theta^0}[\boldsymbol{g};\boldsymbol{g}])^{-1}|V_n|^{-1}\ell_n^{\theta^0}\right] \longrightarrow 0 \quad \text{for } n \to \infty \text{ in probability } [\mu^{\theta^0}].$$

Proof. The statements follow from Proposition 8.1 i) and ii), and the regularity properties of the transform $\theta \mapsto \int f \, \mathrm{d}\mu^\theta$ (cf. Janžura [17] for details). □

Remark 9.2. The preceding theorem states that the MLE $\tilde{\theta}^n$ is asymptotically linearly related to the score function. This yields, together with the regularity ensured by the LAN condition (Proposition 8.1 iii)), the maximum possible concentration about the true value (cf. e.g. Hájek [14] for details).

Proposition 9.3. For every $A \in k(T)$ it holds

$$[B_{\theta^0}^A]^{-1} - [B_{\theta^0}[\boldsymbol{g};\boldsymbol{g}]]^{-1} \geq 0,$$

i.e. the MPLE is asymptotically less efficient to compare with the MLE.

Proof. Since the "asymptotic covariance matrix"

$$\begin{pmatrix} B_{\theta^0}[\boldsymbol{g};\boldsymbol{g}] & -B_{\theta^0}[\boldsymbol{g};S_A^{\theta^0}] \\ -B_{\theta^0}[S_A^{\theta^0};\boldsymbol{g}] & B_{\theta^0}[S_A^{\theta^0};S_A^{\theta^0}] \end{pmatrix}$$

is positive semidefinite, and the particular block submatrices are strictly positive definite by Lemma 5.2 ii) and Lemma 7.3, the statement follows immediately. It also naturally agrees with the well-known Rao–Cramér theorem. Note that $B_{\theta^0}[\boldsymbol{g};\boldsymbol{g}]$ plays the role of the asymptotic Fisher information matrix. □

Remark 9.4. There is an open problem under what reasonable conditions we obtain a strict positive definiteness in the above statement. As a straightforward counterexample we have the i.i.d. case, i.e. $U^1, \ldots, U^N \in \mathcal{U}_0$. Then all the MPL estimates coincide with the MLE, and thus we obtain even the equality of all the asymptotic variance matrices. In general we can observe that the strict positive definiteness occurs whenever the collection

$$U^{\theta^0,A} - B_{\theta^0}\left[S_A^{\theta^0}; \boldsymbol{g}\right] B_{\theta^0}[\boldsymbol{g};\boldsymbol{g}]^{-1} U,$$

where $U = (U^1, \ldots, U^N)^\top$, and for every $i = 1, \ldots, N$ we set $U_B^{\theta^0,A,i} = (S_A^{\theta^0})^i \circ \tau_t$ if $B = \partial A + t$, $t \in T$, and $U_B^{\theta^0,A,i} \equiv 0$ otherwise (cf. also the proof of Lemma 7.3), is given by mutually non-equivalent potentials.

Now, we shall study the behaviour of the asymptotic efficiency for growing A. Let again $\{A_k\}_{k=1}^{\infty}$ satisfy the assumption of Lemma 5.2, namely let $\lim_{k\to\infty} |V_k|^{-1}|A_k \triangle V_k| = 0$.

Theorem 9.5. For $k \to \infty$ the maximum asymptotic efficiency is attained, i.e.

$$\lim_{k\to\infty} B_{\theta^0}^{A_k} = B_{\theta^0}[\boldsymbol{g};\boldsymbol{g}]$$

for every $\theta^0 \in \Theta$.

The proof is given in the following section. □

Remark 9.6. Since $|A_k|^{-1} \left\| F_{A_k} - \sum_{t\in A_k} \boldsymbol{g}\circ\tau_t \right\|_2 \rightrightarrows 0$ uniformly for $k \to \infty$ (cf. the proof of Lemma 10.1 for details) we could replace the term F_{A_k} with $\sum_{t\in A_k} \boldsymbol{g}\circ\tau_t$ in the definition of the function $S_{A_k}^{\theta}$. Setting directly

$$\tilde{S}_k^{\theta} = |A_k|^{-1}\left[E_{A_k}^{\theta}\left[\sum_{t\in A_k} \boldsymbol{g}\circ\tau_k\right] - \sum_{t\in A_k} \boldsymbol{g}\circ\tau_k\right]$$

we can follow this approach from Section 7 to obtain similar results, some of them even in an easier way.

We can make another step and introduce an estimate defined through the function

$$\tilde{\tilde{S}}_{A_k}^{\theta} = E_{A_k}^{\theta}[\boldsymbol{g}] - \boldsymbol{g}.$$

Such kind of estimate obviously strongly imitates the original MLE which can be defined in the same manner only with the "unconditional" $E^{\theta}[\boldsymbol{g}]$. For this particular modifications there would occur some small differences in efficiency, but negligible for $k \to \infty$ since all of them approach the MLE.

In general we can employ any function S^{θ} that ensures regular mapping $\theta \mapsto \int S^{\theta}\,\mathrm{d}\mu$ for every stationary r.f. μ and $0 = \int S^{\theta}\mathrm{d}\mu^{\theta}$ for every $\theta \in \Theta$.

Remark 9.7. The asymptotic results for growing $A_k \nearrow T$ as $k \to \infty$ namely Proposition 6.2 and Theorem 9.5, are naturally not much relevant for the practical purposes since only the MPLE based on a fixed and "rather small" set A can be calculated. Their importance consists in justifying the idea of the MPL estimation as a natural generalization and extension of the ML estimation. We can conclude that by a sequence of the MPL estimates we could approximate the optimum MLE both for a fixed sample size (Proposition 6.2) and in the asymptotic sense (Theorem 9.5.).

10. PROOF OF THEOREM 9.5

For the fixed $\theta^0 = \Phi(U^0) \in \Theta$ we denote $\gamma = \sum_{t \in T} \gamma_t$ where $\gamma_t = \gamma_t(\partial U^0)$ as defined in Section 8. Following Künsch [22] we further denote $\Gamma = (\gamma_{t-s})_{t,s \in T}$ and $\chi = \sum_{n=0}^{\infty} \Gamma^n$. Thus $\chi = (\chi_{ab})_{a,b \in T}$ is an infinite matrix with the property $\sum_{a \in T} \chi_{ab} = \sum_{a \in T} \chi_{ba} = (1-\gamma)^{-1} < \infty$ for every $b \in T$.

Moreover, for a continuous function $f \in C(X^T)$ and $s \in T$ we set

$$\varphi_s(f) = \sup \{|f(x) - f(y)|;\ x_t = y_t \text{ for } t \neq s\}.$$

Let us emphasize that

$$C_W \subset C^1 = \left\{ f \in C(X^T);\ \varphi(f) = \sum_{s \in T} \varphi_s(f) < \infty \right\}$$

for every $W \in k(T)$. Note that $\varphi_s(f \circ \tau_t) = \varphi_{s-t}(f)$ for every $t, s \in T$ and $f \in C^1$.

For some fixed $U \in \mathcal{U}_r$ we again denote

$$g = \sum_{V \in k(T;0)} |V|^{-1} U_V, \quad F_{A_k} = \sum_{V \cap A_k \neq \emptyset} U_V, \quad \text{and} \quad S^{\theta^0}_{A_k} = E^{\theta^0}_{A_k}[F_{A_k}] - F_{A_k}.$$

Lemma 10.1. It holds

i) $|A_k|^{-1} \varphi \left(S^{\theta^0}_{A_k} + \sum_{t \in A_k} g \circ \tau_t \right) \longrightarrow 0$ for $k \to \infty$,

ii) $|A_k|^{-1} \varphi(S^{\theta^0}_{A_k}) \leq$ const. $< \infty$ for every k.

Proof. Let us denote $\mathcal{V}_r = \{V \in k(T);\ \mathrm{diam}(V) \leq r,\ \min V = 0\}$, where the minimum is taken with respect to some linear (e.g. the lexicographical) ordering. Note that $|\mathcal{V}_r| < \infty$ and $\mathcal{V}_r$ does not contain any pair of "shift-similar" sets.

Further, denoting $A_k \ominus V = \{t \in T;\ (V + t) \cap A_k \neq 0\}$ for $V \in \mathcal{V}_r$ we observe

$$|A_k \ominus V| \leq |A_k| \cdot |V|,$$

and for fixed $b \in T$ we obtain

$$|A_k|^{-1} |(A_k \ominus V) \cap (A_k^c - b)| \longrightarrow 0 \quad \text{for } k \to \infty.$$

Now, since $F_{A_k} = \sum_{V \in \mathcal{V}_r} \sum_{t \in A \ominus V} U_{V+t}$ and $\sum_{t \in A_k} g \circ \tau_t = \sum_{V \in \mathcal{V}_r} \sum_{t \in A \ominus V} \frac{U_{V+t}}{|V|} |(U+t) \cap A_k|$ we obtain

$$|A_k|^{-1} \varphi \left(F_{A_k} - \sum_{t \in A_k} g \circ \tau_t \right) \leq \sum_{V \in \mathcal{V}_r} \varphi(U_V) \cdot |A_k|^{-1} \sum_{t \in A_k \ominus V} \left(1 - \frac{|(A_k - t) \cap V|}{|V|} \right)$$

$$= \sum_{V \in \mathcal{V}_r} \varphi(U_V) \, |A_k|^{-1} |V|^{-1} \sum_{s \in V} |(A_k \ominus V) \cap (A_k^c - s)| \longrightarrow 0 \quad \text{for } k \to \infty.$$

Further, by Corollary 2.4 in Künsch [22] we obtain for $b \notin A_k$

$$\varphi_b \left(E_{A_k}^{\theta^0} [F_{A_k}] \right) \leq \sum_{q \in T} \chi_{bq} \varphi_q(F_{A_k}),$$

while for $b \in A_k$ we have zero by definition, and therefore

$$|A_k|^{-1} \varphi \left(E_{A_k}^{\theta^0} [F_{A_k}] \right) \leq |A_k|^{-1} \sum_{b \notin A_k} \sum_{q \in T} \chi_{bq} \varphi_b(F_{A_k})$$

$$\leq \sum_{V \in \mathcal{V}_r} |A_k|^{-1} \sum_{b \notin A_k} \sum_{t \in A_k \ominus V} \sum_{q \in T} \chi_{bq} \varphi_{q-t}(U_V)$$

$$= \sum_{V \in \mathcal{V}_r} \sum_{b,q \in T} \chi_{bq} \varphi_q(U_V) |A_k|^{-1} |(A_k \ominus V) \cap (A_k^c - b)| \longrightarrow 0 \quad \text{for } k \to \infty$$

by the dominated convergence arguments since $|A_k|^{-1} |(A_k \ominus V) \cap (A_k^c - b)| \leq |V|$. Hence, since $\varphi \left(S_{A_k}^{\theta^0} + \sum_{t \in A_k} g \circ \tau_t \right) \leq \varphi(F_{A_k} - \sum_{t \in A_k} g \circ \tau_t) + \varphi \left(E_{A_k}^{\theta^0}(F_{A_k}) \right)$ the proof of i) is completed.

Further, since $\varphi \left(E_{A_k}^{\theta^0}(F_{A_k}) \right) \leq (1-\gamma)^{-1} \varphi(F_{A_k})$ and $|A_k|^{-1} \varphi(U_V) \, |A_k \ominus V| \, |A_k|^{-1} \leq \sum_{V \in \mathcal{V}_r} |V| \, \varphi(U_V)$, the proof of ii) is straightforward. □

Now for U^i, $i = 1, \ldots, N$, we shall denote by g^i and $[S_{A_k}^{\theta^0}]^i$, respectively, the corresponding terms.

Proposition 10.2. For every pair (i, j) it holds

$$|A_k|^{-1} B_{\theta^0} \left(g^i, [S_{A_k}^{\theta^0}]^j \right) + B(g^i, g^j) \longrightarrow 0 \quad \text{for } k \to \infty$$

and

$$|A_k|^{-2} B_{\theta^0} \left([S_{A_k}^{\theta^0}]^i, [S_{A_k}^{\theta^0}]^j \right) + |A_k|^{-1} B_{\theta^0} \left(g^i, [S_{A_k}^{\theta^0}]^j \right) \longrightarrow 0 \quad \text{for } k \to \infty.$$

Proof. Since obviously

$$B_{\theta^0}(g^i, g^j) = |A_k|^{-1} B_{\theta^0} \left(g^i, \sum_{t \in A_k} g^j \circ \tau_t \right)$$

we may write

$$\begin{aligned}
&\left||A_k|^{-1}B_{\theta^0}\left(g^i,[S_{A_k}^{\theta^0}]^j\right)+B_{\theta^0}(g^i,g^j)\right| \\
&= |A_k|^{-1}\left|B_{\theta^0}\left(g^i,[S_{A_k}^{\theta^0}]^j+\sum_{t\in A_k}g^j\circ\tau_t\right)\right| \\
&\leq |A_k|^{-1}\sum_{s\in T}\sum_{a,b,c\in T}\sum\chi_{ca}\chi_{cb}\varphi_a(g^i\circ\tau_s)\,\varphi_b\left([S_{A_k}^{\theta^0}]^j+\sum_{t\in A_k}g^j\circ\tau_t\right) \\
&= (1-\gamma)^{-2}\,\varphi(g^i)\,|A_k|^{-1}\,\varphi\left([S_{A_k}^{\theta^0}]^j+\sum_{t\in A_k}g^j\circ\tau_t\right)
\end{aligned}$$

by Corollary 3.4 in Künsch [22].

Similarly

$$\begin{aligned}
&\left||A_k|^{-2}B_{\theta^0}\left([S_{A_k}^{\theta^0}]^i,[S_{A_k}^{\theta^0}]^j\right)+|A_k|^{-1}B_{\theta^0}\left(g^i,[S_{A_k}^{\theta^0}]^j\right)\right| \\
&\leq (1-\gamma)^{-2}|A_k|^{-1}\,\varphi\left([S_{A_k}^{\theta^0}]^j\right)|A_k|^{-1}\,\varphi\left([S_{A_k}^{\theta^0}]^i+\sum_{t\in A_k}g^i\circ\tau_t\right).
\end{aligned}$$

Thus both the terms tend to zero by Lemma 10.1. □

Corollary 10.3. It holds

$$\lim_{k\to\infty}B_{\theta^0}^{A_k}=B_{\theta^0}[\boldsymbol{g};\boldsymbol{g}].$$

Proof. Since all the involved matrices are positive definite, the assertion follows from the "term-wise" convergence ensured by Proposition 10.2. □

11. INFINITESIMAL ROBUSTNESS

Finally, we shall briefly discuss the problem of robustness, which is understood as a sensitiveness of the estimator to the data.

For fixed $\theta^0\in\Theta$ and $A\in k(T)$ we set

$$\theta_A(\varepsilon,\nu)=\left[J^A_{(1-\varepsilon)\mu^{\theta^0}+\varepsilon\nu}\right]^{-1}(0)$$

i.e.

$$\int S_A^{\theta_A(\varepsilon,\nu)}d\left[(1-\varepsilon)\,\mu^{\theta^0}+\varepsilon\nu\right]=0$$

for every stationary r.f. ν and $\varepsilon>0$. By direct differentiating and proper substituting we obtain

$$\theta'_A(0,\nu)=\left.\frac{\mathrm{d}\theta_A(\varepsilon,\nu)}{\mathrm{d}\varepsilon}\right|_{\varepsilon=0}=B_{\theta^0}\left[\boldsymbol{g};S_A^{\theta^0}\right]^{-1}\int S_A^{\theta^0}\,\mathrm{d}\nu.$$

Since for every ν and large enough n we have $\int S_A^{\theta^0} \mathrm{d}\nu = \int \left[\int S_A^{\theta^0} \mathrm{d}\tilde{\mu}_{\hat{x}}^n\right] \mathrm{d}\nu(\hat{x})$, we can understand in general

$$\theta'_A(0, \tilde{\mu}_{\hat{x}}^n) = B_{\theta^0}\left[\boldsymbol{g}; S_A^{\theta^0}\right]^{-1} \int S_A^{\theta^0} \mathrm{d}\tilde{\mu}_{\hat{x}}^n$$

as the influence function for the MPLE $\tilde{\theta}_A^n$, i.e.

$$\theta_A(\varepsilon, \nu) \approx \theta^0 + \varepsilon \int \theta'_A(0, \tilde{\mu}_{\hat{x}}^n)\, \mathrm{d}\nu(\hat{x})$$

for "small" $\varepsilon > 0$, and for ν "not too far" from μ^{θ^0} we obtain

$$\theta_A(1, \nu) \approx \theta^0 + \int \theta'_A(0, \tilde{\mu}_{\hat{x}}^n)\, \mathrm{d}\nu(\hat{x}).$$

Analogously from Remark 7.2, for the MLE $\tilde{\theta}^n$ we obtain the influence function

$$\theta'(0, \tilde{\mu}_{\hat{x}}^n) = B_{\theta^0}[\boldsymbol{g}; \boldsymbol{g}]^{-1} \int \left[\boldsymbol{g} - \int \boldsymbol{g}\, \mathrm{d}\mu^{\theta^0}\right] \mathrm{d}\tilde{\mu}_{\hat{x}}^n.$$

Proposition 11.1. The influence functions $\theta'_A(0, \tilde{\mu}_{\hat{x}}^n)$ and $\theta'(0, \tilde{\mu}_{\hat{x}}^n)$, respectively, are zero mean random variables under μ^{θ^0}, satisfying

$$\theta'_A(0, \tilde{\mu}_{\hat{x}}^n) \longrightarrow 0, \quad \theta'(0, \tilde{\mu}_{\hat{x}}^n) \longrightarrow 0 \quad \text{a.s. } [\mu^{\theta^0}] \text{ for } n \to \infty,$$

and

$$|V_n|^{\frac{1}{2}}\, \theta'_A(0, \tilde{\mu}_{\hat{x}}^n) \Longrightarrow \mathcal{N}_N(0, [B_{\theta^0}^A]^{-1}), \quad |V_n|^{\frac{1}{2}}\, \theta'(0, \tilde{\mu}_{\hat{x}}^n) \Longrightarrow \mathcal{N}_N(0, B_{\theta^0}[\boldsymbol{g}; \boldsymbol{g}]^{-1})$$
$$\text{in distribution } [\mu^{\theta^0}] \text{ for } n \to \infty.$$

Proof. The statements follow from the properties of the unbiased version empirical r.f. $\tilde{\mu}_{\hat{x}}^n$, ergodicity of μ^{θ^0}, and Proposition 8.1.i). □

Corollary 11.2. The every MPLE $\tilde{\theta}_A^n$ is "less robust" than the MLE $\hat{\theta}^n$ since $[B_{\theta^0}^A]^{-1} - [B_{\theta^0}[\boldsymbol{g}, \boldsymbol{g}]]^{-1} \geq 0$, i.e. under the true model μ^{θ^0} the asymptotic covariance matrix of the influence function $\theta'_A(0, \tilde{\mu}_{\hat{x}}^n)$ is "greater" than the asymptotic covariance matrix of $\theta'(0, \tilde{\mu}_{\hat{x}}^n)$,

Proof. The statement follows directly from the preceding proposition and Remark 9.4. □

Note that the above results are rather natural since all the considered estimators belong to the class of the so-called M-estimators with bounded ψ-function which are, also from the robustness point of view, deeply studied and well understood in mathematical statistics. For general treating cf. e.g. Hampel et al [15], and for the case of autoregressive processes cf. Künsch [23].

ACKNOWLEDGEMENT

The author wishes to express his gratitude to Gábor Tusnádi who formulated the problem and conjectured the results of Theorem 9.5.

(Received November 3, 1995.)

REFERENCES

[1] J. Besag: Spatial interaction and the statistical analysis of lattice systems. J. Roy. Statist. Soc. B *36* (1974), 192–226.
[2] J. Besag: On the statistical analysis of dirty pictures (with discussion). J. Roy. Statist. Soc. Ser. B *48* (1986), 259–302.
[3] E. Bolthausen: On the central limit theorem for stationary mixing random fields. Ann. Probab. *10* (1982), 1047–1050.
[4] F. Comets: On consistency of a class of estimators for exponential families of Markov random fields on a lattice. Ann. Statist. *20* (1992), 455–468.
[5] R. L. Dobrushin and B. S. Nahapetian: Strong convexity of the pressure for the lattice systems of classical statistical physics. Teor. Mat. Phys. *20* (1974), 223–234.
[6] D. Geman and S. Geman: Maximum Entropy and Bayesian Methods in Sciences and Engineering (C. R. Smith and G. J. Erickson, eds.), Kluwer, Dordrecht 1988.
[7] S. Geman and C. Graffigne: Markov random field image models and their applications to computer vision. In: Proc. Internat. Congress Math. (A. M. Gleason ed.), Amer. Math. Soc., Providence, R. I. 1987.
[8] H. O. Georgii: Gibbs Measures and Phase Transitions. De Gruyter, Berlin 1988.
[9] B. Gidas: Consistency of maximum likelihood and pseudo–likelihood estimators for Gibbs distribution. In: Stochastic Differential Systems, Stochastic Control Theory, and Application (W. Fleming and P. L. Lions, eds., IMA Vol. Math. Appl. 10). Springer, New York 1988.
[10] B. Gidas: Parameter estimation for Gibbs distributions. I. Fully observed data. In: Markov Random Fields: Theory and Applications (R. Chellapa and R. Jain, eds.), Academic Press, New York 1991.
[11] L. Gross: Absence of second-order phase transition in the Dobrushin's uniqueness region. J. Statist. Phys. *27* (1981), 57–72.
[12] X. Guyon: Estimation d'un champ par pseudo-vraisemblance conditionnelle: Etude asymptotique et application au cas Markovien. In: Actes de la 6^{eme} Rencontre Franco–Belge de Statisticiens, Bruxelles 1985.
[13] X. Guyon and H. R. Künsch: Asymptotic comparison of estimators in the Ising model. In: Stochastic Models, Statistical Methods, and Algorithms in Image Analysis (P. Barone, A. Frigessi and M. Piccioni, eds., Lecture Notes in Statistics 74), Springer, Berlin 1992, pp. 177–198.
[14] J. Hájek: Local asymptotic minimax and admissibility in estimation. In: Proc. 6th Berkeley Symposium, Vol. 1, Berkeley, Calif. 1970, pp. 175–194.
[15] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw and W. A. Stahel: Robust Statistics – The Approach Based on Influence Functions. Wiley, New York 1986.
[16] M. Janžura: Estimating interactions in binary lattice data with nearest–neighbor property. Kybernetika *23* (1987), 2, 136–142.
[17] M. Janžura: Statistical analysis of Gibbs random fields. In: Trans. 10th Prague Conf. on Inform. Theory, Stat. Dec. Functions, Random Processes 1986, Praha, pp. 429–438.
[18] M. Janžura: Asymptotic theory of parameter estimation for Gauss–Markov random fields. Kybernetika *24* (1988), 161–176.

[19] M. Janžura: Local asymptotic normality for Gibbs random fields. In: Proceedings of the Fourth Prague Symposium on Asymptotic Statistics (P. Mandl, M. Hušková, eds.), Charles University, Prague 1989, pp. 275–284.
[20] M. Janžura and P. Lachout: A central limit theorem for stationary random fields. Math. Methods Statist. *4* (1995), 463–472.
[21] H. R. Künsch: Thermodynamics and statistical analysis of Gaussian random fields. Z. Wahrsch. Verw. Gebiete *58* (1981), 407–421.
[22] H. Künsch: Decay of correlations under Dobrushin's uniqueness condition and its applications. Commun. Math. Phys. *84* (1982), 207–222.
[23] H. Künsch: Infinitesimal robustness for autoregressive processes. Ann. Statist. *12* (1984), 843–863.
[24] C. Preston: Random Fields (Lecture Notes in Mathematics 534). Springer, Berlin 1976.
[25] D. J. Strauss: Analysing binary lattice data with the nearest–neighbor property. J. Appl. Probab. *12* (1975), 702–712.
[26] L. Younès: Estimation and annealing for Gibbsian fields. Ann. Inst. H. Poincaré Sect. B (N.S.) *24* (1988), 269–294.
[27] L. Younès: Parametric inference for imperfectly observed Gibbsian fields. Probab. Theory Related Fields *82* (1989), 625–645.

RNDr. Martin Janžura, CSc., Ústav teorie informace a automatizace AV ČR (Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic), P d vodárenskou věží 4, 182 08 Praha 8. Czech Republic.