SIEVE ESTIMATORS FOR PAIR-INTERACTION POTENTIALS AND LOCAL CHARACTERISTICS IN GIBBS RANDOM FIELDS

CHUANSHU JI

Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260
USA

Summary

The problems of estimating certain infinite-dimensional unknown parameters (such as interaction potentials, local characteristics, etc.) in Gibbs random fields are considered. We apply Grenander's method of sieves to construct strongly consistent estimators for those parameters. Exponential rates of consistency have been established by using the conditional mixing property of the Gibbs random fields. The approach of the paper is applicable to image models with pair-potentials of short range or long range and with simple degradation structure, such as texture segmentation models. The results in the paper hold regardless of phase transition and symmetry breaking in the Gibbs random fields.

Key words and phrases: Gibbs random fields, method of sieves, consistent estimator, texture segmentation, image analysis.

Abbreviated Title: Estimation for Gibbs Random Fields.

AMS (1980) subject classification:
primary 62G05, 60G60; secondary 68G10, 82A25, 82A67.

1. Introduction.

Since the pioneer work of Geman and Geman (1984), Gibbs random fields (GRF) have been extensively used in imaging problems. GRF were originally introduced as models in statistical mechanics. A particle on each site of a two-dimensional integer lattice may represent the "spins" of a magnet, and there are usually interactions between particles at different sites. A configuration consists of particles on all sites. GRF are probability distributions on the set of all configurations. In imaging problems, particles are replaced by picture elements ("*pixels*"). Intrinsic properties of GRF are preserved. In terms of Bayesian paradigm, GRF play a role of the prior distribution on relevant scene attributes to capture the tendencies and constraints that characterize the scene of interest. Image processing is then guided by the prior, which, if properly conceived, enormously limits the plausible restorations and interpretations. The mean or mode(s) of the prior, if regarded as the true scene(s), usually can only be inferred based upon partial or corrupted observations, i.e. the corresponding posterior mean or posterior mode(s) are taken as the estimates of the true scene. For instance, in emission tomography, the true scene is represented by the spatial distribution of isotope in a target region of the body. The observations - photon counts - whose probability law is specified by the likelihood function given the true scene, have a mean function that is an attenuated Radon transform of the isotope intensity. In the texture segmentation problems, an image consists of the pixel intensity array and a corresponding array of texture lables. Each label gives the texture type of the associated pixel. The grey-levels of the pixels are observed, but not the labels. There is a substantial literature on imaging problems from the viewpoint of probabilistists and statisticians. We refer the reader to Geman and Geman (1984), Besag (1986) and Geman (1990) for more complete discussions of general framework and

methodology, Geman and McClure (1987), Johnstone and Silverman (1990) for tomography, Geman and Graffigne (1986) for texture segmentation.

The quality of image processing will clearly depend on choices made at the modeling stage, i.e. how to specify GRF as a prior. This may include two aspects: *parameter estimation* and *model selection*. The former is to estimate unknown parameters contained in the energy functions of GRF. The latter is to choose one from several candidate GRF as the true model. In this paper we only consider parameter estimation. Model selection will be discussed later on in Seymour and Ji (1990).

The parameter estimation for GRF is complicated by *phase transition* and *degradation* that creates corrupted scenes as indirect observations. Phase transition means that for certain values of the true parameters there may exist more than one "infinite-volume" GRF that generate the data. It produces an interesting class of spatial statistical models with long-range dependence, which can not be detected solely by the data in advance. Therefore good estimators should have nice asymptotic properties which hold regardless of phase transitions, and meanwhile be computationally feasible.

So far all available estimation procedures have been parametric, i.e. they are based on the assumption that the energy functions in the GRF are parametrized by a finite number of unknown parameters. These procedures include *maximum likelihood esimators* (MLE), *maximum pseudo-likelihood estimators* (MPLE), etc. Computationally, MLE are intractable in their basic forms. Some work has been done to remedy this drawback. Meanwhile, MPLE turn out to be very effective. Gidas (1988) established the consistency, asymptotic normality and asymptotic efficiency of MLE for fully observed data. Cométs and Gidas (1988) proved the consistency of MLE for degraded data. For MPLE, Geman and Graffigne (1986), Gidas (1986) obtained the consistency respectively by using different approaches.

Recently, Cométs (1989) applied the large deviation theory to derive the exponential convergence rates for the consistency of MLE and MPLE, and also studied the connection with the Bahadur efficiency. It is noteworthy that the consistency of those estimators is not affected by phase transition under some identifiability conditions, but the asymptotic normality is.

As mentioned in Gidas (1988) and Geman (1990), an open problem along this line is nonparametric estimation in which the unknown parameters contained in the energy functions of the GRF are infinite-dimensional: either infinite sequences or smooth functions. Why should we consider infinite-dimensional parameters? Mainly because it may be a good starting point when we know very little about the energy functions. In this paper, Grenander's method of *sieves* [Grenander (1981), Geman and Hwang (1982)] is adopted to construct strongly consistent estimators for the local characteristics (as unknown functions) in GRF, the results then are used to produce strongly consistent estimates for a countable sequence of unknown coefficients of the pair-potentials in the energy functions. Consistency of MPLE is also proved as a by-product for the case in which the energy functions contain a large number of parameters. This generalizes some results in the papers aforementioned. The main ingredient of sieve method in our context is to choose an increasing sequence of sieves - each being a constrained finite-dimensional subspace of the original infinite-dimensional parameter space - so that every sieve implicitly corresponds to a Markov random field (MRF) induced by interactions of finite range. As the sample size increases, so is the sieve size, but with a slower rate. The relation between the growth rate of sieve size and the convergence rate in the consistency of our sieve estimators is carefully demonstrated, which indicates how far this method can go.

This paper may be the first one to make an attempt on nonparametric estimation for GRF. The results hold regardless of phase transition and *symmetry breaking* (GRF need not be stationary). The estimators we constructed are quite

tractable computationally. However, a few limitations should be mentioned here.

(i) The sieve estimators in this paper apply only to the cases of fully observed data or the incomplete data such as grey-levels in texture segmentation problems, where the degradation is simply a projection and the parameters of interest are only related to the grey-levels, so it can be dealt with as if there is no degradation. Nonparametric estimation problems for the GRF with more complex degradation structure are yet to be tackled.

(ii) For the data set consisting of observations taken from an n×n square lattice, we choose an k×k square lattice as a sieve compatible with the sample size, where k behaves like $\sqrt{c \log n}$ for large n. To determine the constant c, which depends on the unknown parameters, a bound for the sum of the external field coefficient and the pair-potentials is assumed to be known as a priori. This is similar to the situation in density estimation, where the bandwidth selection is based on the knowledge of bound on the derivative of the density to be estimated. Therefore, effort in the future should be made to replace this assumption by sensible adaptive procedures to determine c from the data.

In Section 2 we give the background of GRF and formulate the related estimation problems. The main results are stated in Section 3. A heuristic argument for choosing the sieve size is also given there. In Section 4, we derive the rate of consistency for the sieve estimators of the local characteristics, using the conditional mixing properties of the GRF and a lower bound on the probabilities of possible configurations restricted to each sieve. In Section 5, we deal with the consistent estimators for pair-potentials in GRF of certain types, including the general Ising model and the one introduced by Geman and Graffigne (1986) for texture segmentation. Generalization of the consistency for MPLE is discussed. Finally, in Section 6, we make some concluding remarks and mention some possible issues for future research.

The term "*nonparametric*" is usually used for "*model-free*" statistics. However, using GRF to build up image models tends to comply with "*model-based*" principle. In a sense the estimation procedures in this paper could be called "*infinite-parametric*". Nevertheless, following the tradition, we still adopt "nonparametric" here.

2. Mathematical model and preliminaries.

2.1 GRF induced by pair-interaction potentials.

Let $\mathbb{Z}^2$ be the two-dimensional integer lattice. With each pixel site $\xi \in \mathbb{Z}^2$, we associate a random variable X_ξ taking values in a finite set $S = \{\pm 1, \ldots, \pm r\}$, where $r \in \mathbb{N}$. Let $\Omega = S^{\mathbb{Z}^2}$ be the configuration space and each $x \in \Omega$ be a realization of X. Here $x = (x_\xi,\ \xi \in \mathbb{Z}^2)$, $X = (X_\xi,\ \xi \in \mathbb{Z}^2)$. For every $\Lambda \subset \mathbb{Z}^2$, we let $\Omega_\Lambda = S^\Lambda$, and $x_\Lambda = (x_\xi,\ \xi \in \Lambda) \in \Omega_\Lambda$, $X_\Lambda = (X_\xi,\ \xi \in \Lambda)$. In imaging problems, x_ξ may represent *the grey-level* of the pixel ξ.

GRF will be the probability distributions of X defined on Ω via *pair-interaction potentials* (or simply pair-potentials in what follows).

Let $J = \{J(\xi),\ \xi \in \mathbb{Z}^2\}$ be a real sequence satisfying

(A1) $J(\xi) = J(-\xi)$, $\forall\ \xi \in \mathbb{Z}^2$, and $J(\omega) = 0$, ω is the origin of $\mathbb{Z}^2$.

(A2) $\exists\ C' > 0$, $a > 0$ such that

$$|J(\xi)| \le C'\ e^{-a\|\xi\|},\quad \text{uniformly } \forall\ \xi \in \mathbb{Z}^2,$$

where the norm $\|\cdot\|$ on $\mathbb{Z}^2$ is defined by $\|\xi\| = \max\{|\xi_1|,\ |\xi_2|\}$ for $\xi = (\xi_1, \xi_2)$.

Let $V : S \to \mathbb{R}$ and $U : S \times S \to \mathbb{R}$ be two functions satisfying

(A3) V is either identically zero or nonconstant. Moreover, U is symmetric, and for every $\tilde{z} \in S$

$$\max_{z \in S} U(\tilde{z}, z) > \min_{z \in S} U(\tilde{z}, z).$$

(A4) $\exists\ z_1,\ z_2 \in S$, called *double extremal points*, such that one of the following two cases holds:

Case 1: $V \equiv 0$; $U(z_1,\bar{z}) = \max_{z\in S} U(z,\bar{z})$, $U(z_1,\underline{z}) = \min_{z\in S} U(z,\underline{z})$ for some $\bar{z},\ \underline{z} \in S$.

Case 2: V is nonconstant, $V(z_1) = \max_{z\in S} V(z)$, $V(z_2) = \min_{z\in S} V(z)$; meanwhile,

$$U(z_1,\bar{z}) = \max_{z\in S} U(z,\bar{z}),\ U(z_1,\underline{z}) = \min_{z\in S} U(z,\underline{z}),\ \text{and}\ U(z_2,z^*) = \max_{z\in S} U(z,z^*),$$

$$U(z_2,z_*) = \min_{z\in S} U(z,z_*) \text{ for some } \bar{z},\ \underline{z},\ z^*,\ z_* \in S.$$

The pair-potential between the pixels ζ and η is assumed to have the form $J(\zeta-\eta)\ U(x_\zeta,x_\eta)$. Apparently, such potentials are translation-invariant. They also generalize the potentials of finite range in which $J(\xi) = 0$ for all ξ with sufficiently large norm $\|\xi\|$.

For every finite subset Λ of $\mathbb{Z}^2$ and each $x \in \Omega$, define *the energy*

$$(2.1)\quad H_\Lambda(x) = -h \sum_{\xi\in\Lambda} V(x_\xi) - \frac{1}{2} \sum_{\zeta,\eta\in\Lambda} J(\zeta-\eta)\ U(x_\zeta,x_\eta) - \sum_{\zeta\in\Lambda} \sum_{\eta\in\Lambda^c} J(\zeta-\eta)\ U(x_\zeta,x_\eta),$$

where $h \in \mathbb{R}$ is called *the coefficient of external field*. $H_\Lambda(x)$ may be interpreted as the contribution of the pixels in Λ to the total energy associated with the configuration x.

The finite-volume Gibbs distribution in the volume Λ with the external condition x_{Λ^c} is given by

$$(2.2)\qquad f_\Lambda(y_\Lambda | x_{\Lambda^c}) = \frac{e^{-H_\Lambda(y_\Lambda \oplus x_{\Lambda^c})}}{Z_{\Lambda,x_{\Lambda^c}}},\quad \forall\ y_\Lambda \in \Omega_\Lambda,$$

where the combined configuration $y_\Lambda \oplus x_{\Lambda^c}$ agrees with y_Λ in Λ, and x_{Λ^c} in Λ^c; the normalizing factor

$$(2.3)\qquad Z_{\Lambda,x_{\Lambda^c}} = \sum_{y_\Lambda\in\Omega_\Lambda} e^{-H_\Lambda(y_\Lambda\oplus x_{\Lambda^c})}$$

is called *the partition function* on Λ given x_{Λ^c}.

In particular, when $\Lambda = \{\xi\}$ (a singleton), let ${}_{\xi}x = x_{\{\xi\}^c}$, $f_{\xi}(\cdot|\cdot) = f_{\{\xi\}}(\cdot|\cdot)$, then for a given ${}_{\xi}x$,

$$f_{\xi}(y_{\xi}|{}_{\xi}x) = Z^{-1}_{\{\xi\},{}_{\xi}x}\, e^{-H_{\{\xi\}}(y_{\xi}\oplus_{\xi}x)}, \quad y_{\xi} \in S \tag{2.4}$$

is called *the local characteristics* at ξ. It is known that the local characteristics in (2.4) determine the finite-volume Gibbs distributions in (2.2).

Example 2.1. Let $V(z) = z$, $U(z,\tilde{z}) = z\tilde{z}$, $z,\tilde{z} \in S$. This corresponds to *the general Ising models* (GIM). The double extremal points $z_1 = r$ with $\bar{z} = r$ and $\underline{z} = -r$, while $z_2 = -r$ with $z^* = -r$ and $z_* = r$ in (A4). This is Case 2.

Example 2.2. Let $V(z) = 0$, $\forall\, z \in S$, and $U(z,\tilde{z}) = \dfrac{1}{1+\sigma(z-\tilde{z})^2}$, $z,\tilde{z} \in S$, where σ is a positive constant. The pair-potentials are connected with the texture segmentation models described in Geman and Graffigne (1986). The existence of double extremal point z_1, can also be verified: for instance, $z_1 = r$ with $\bar{z} = r$ and $\underline{z} = -r$ in (A4). This is Case 1.

For the potentials specified by U, V, J, h, let $\mathcal{G}(U,V;J,h) \triangleq \mathcal{G}$ be the set of corresponding *infinite-volume* GRF on Ω so that $P \in \mathcal{G}$ if for each finite $\Lambda \subset \mathbb{Z}^2$ and every $x_{\Lambda^c} \in \Omega_{\Lambda^c}$,

$$P(X_{\Lambda} = y_{\Lambda} \mid X_{\Lambda^c} = x_{\Lambda^c}) = f_{\Lambda}(y_{\Lambda}|x_{\Lambda^c}), \quad \forall\, y_{\Lambda} \in \Omega_{\Lambda}. \tag{2.5}$$

P is said to be stationary if

$$P(X_{\Lambda+\xi} = x_{\Lambda}) = P(X_{\Lambda} = x_{\Lambda}), \quad \forall \text{ finite } \Lambda \subset \mathbb{Z}^2,\ \xi \in \mathbb{Z}^2,\ x_{\Lambda} \in \Omega_{\Lambda}, \tag{2.6}$$

where $X_{\Lambda+\xi} = (X_{\zeta+\xi},\ \zeta \in \Lambda)$. Note that the translation invariant potentials need not induce stationary GRF (possibility of symmetry breaking).

Meanwhile, under our assumption, $\mathcal{G}$ is always non-empty, but need not be a singleton (possibility of phase transition). In general, $\mathcal{G}$ is a convex, compact Choquet simplex.

Remark. The definition of GRF under more general conditions is given in Ruelle (1978) and Georgii (1988). Both the configuration space and the potentials can be brought into a more general set-up. However, the framework given here is convenient for the nonparametric estimation problems related to GRF.

2.2 The nonparametric estimation problems for GRF.

Suppose h and J are unknown parameters which induce the set of GRF $\mathcal{G}$ on Ω. Let Λ_n be the n×n symmetric square lattice centered at the origin ω of $\mathbb{Z}^2$, here without loss of generality we assume $n \in \mathbb{N}$ is odd. Based on the data $X(n) \triangleq X_{\Lambda_n}$ generated by a GRF $P \in \mathcal{G}$, two nonparametric estimation problems are considered:

(I) estimating the local characteristics $f_\omega(x)$, $x \in \Omega$, where $f_\omega(x) = f_\omega(x_\omega|_\omega x)$;

(II) estimating h and J.

Note that the translation invariance of the potentials implies the translation invariance of the local characteristics, i.e.

$$f_\xi(t^\xi x) = f_\omega(x), \quad \forall\, \xi \in \mathbb{Z}^2, \quad x \in \Omega, \tag{2.7}$$

where t^ξ is the translation operator defined by $(t^\xi x)_\zeta = x_{\zeta-\xi}$, $\forall\, \zeta \in \mathbb{Z}^2$. So (I) amounts to estimating $f_\xi(x_\xi|_\xi x)$ $\forall\, \xi \in \mathbb{Z}^2$.

A random function T_n on Ω constructed from X(n) is said to be *a strongly consistent estimator of* f_ω if

$$\|T_n - f_\omega\| \to 0, \quad \text{a.s. under } P \text{ as } n \to \infty, \tag{2.8}$$

where $\|T_n - f_\omega\| \triangleq \sup_{x\in\Omega} |T_n(x) - f_\omega(x)|$.

A random sequence $(\hat{h}_n;\ \hat{J}_n(\xi),\ \xi \in \mathbb{Z}^2)$ is called *a strongly consistent estimator of* (h;J) if

$$(2.9) \qquad |\hat{h}_n - h| + \sum_{\xi \in \mathbb{Z}^2} |\hat{J}_n(\xi) - J(\xi)| \to 0, \quad \text{a.s.}$$

under P as $n \to \infty$.

Here we adopt the ℓ^1-distance because (A2) implies the summability of J.

There is another estimation problem which is closely related to (II) but practically more appealing. Suppose X(n) is generated by a MRF with respect to a set of pair-potentials of a finite but large range. i.e. the energy function contains finite but a large number of unknown parameters. To formulate this quantitatively, we may assume that the range of the potentials, or equivalently the number of unknown parameters is d_n -- it increases along with the sample size but with a slower rate. This is essentially a nonparametric estimation problem even though it looks like a parametric model superficially. Intuitively, the question here is given a large data set what would be our limitation in terms of the compatible number of unknown parameters which we can estimate consistently. We call the problem

(III) estimating h and $J = \{J(\xi) : \xi \in \mathbb{Z}^2, \|\xi\| \le d_n\}$.

It should be noticed that we do not consider the problem of estimating the GRF themselves here. Due to the phase transitions, the GRF themselves are generally non-identifiable. See Gidas (1988) for further discussions.

3. Main results.

3.1. Construction of T_n and selection of the sieve size k.

We write f(x) for $f_\omega(x)$, $x \in \Omega$ in the problem (I) of Section 2.2 when there is no confusion. By (2.5), f(x) is the conditional probability $P(X_\omega = x_\omega | {}_\omega X = {}_\omega x)$, which can naturally be approximated by

$$(3.1) \qquad f^{(k)}(x) \triangleq P(X_\omega = x_\omega | X_{\Lambda_k \setminus \{\omega\}} = x_{\Lambda_k \setminus \{\omega\}})$$

for large $k \in \mathbb{N}$, where Λ_k -- the sieve -- is an k×k square lattice centered at ω. Therefore we construct the empirical measure from the sample X(n), and use the "*sample conditional frequency*" of x_ω appearing at ω given that $x_{\Lambda_k \setminus \{\omega\}}$ appearing in $\Lambda_k \setminus \{\omega\}$ to estimate $f^{(k)}(x)$. More specifically, we first extend X(n) by periodization outside Λ_n into a periodic configuration X^n, then define

$$R_{n,X} = \frac{1}{|\Lambda_n|} \sum_{\xi \in \Lambda_n} \delta_{t^\xi X^n}. \tag{3.2}$$

where δ_x is the Dirac mass at $x \in \Omega$; $|\Lambda|$ is the cardinality of $\Lambda \subset \mathbb{Z}^2$. $R_{n,X}$ is the empirical measure defined on Ω constructed from X(n).

For every $x_{\Lambda_k} \in \Lambda_{\Lambda_k}$, let

$$A_k = \{y \in \Omega : y_{\Lambda_k} = x_{\Lambda_k}\} ; \tag{3.3}$$

$$A_k' = \{y \in \Omega : y_{\Lambda_k \setminus \{\omega\}} = x_{\Lambda_k \setminus \{\omega\}}\} . \tag{3.4}$$

Define $T_n : \Omega \to \mathbb{R}$ by

$$T_n(x) = \begin{cases} R_{n,X}(A_k)/R_{n,X}(A_k'), & \text{if } R_{n,X}(A_k) > 0, \\ \tilde{a} & , \text{ otherwise,} \end{cases} \tag{3.5}$$

where $\tilde{a} \in (0,1)$ can be set arbitrarily whose value is not important. T_n, as the "sample conditional frequency" mentioned before, is our estimator for f.

Now we give a heuristic argument that for large n, the sieve size k should behave like $\sqrt{c \log n}$ to guarantee the consistency of T_n.

3.1.1 Why $\sqrt{\log n}$?

Suppose X(n) comes from an extremal point $P \in \mathcal{G}$ so X is ergodic under P. Let

$$P(x(k)) = P(X_{\Lambda_k} = x_{\Lambda_k}), \quad x_{\Lambda_k} \in \Omega_{\Lambda_k} . \tag{3.6}$$

By the Shannon-McMillan-Breiman Theorem [cf. Föllmer (1973)], for P-almost all

$x \in \Omega$,

$$P(x(k)) \approx e^{-|\Lambda_k| h(P)}, \quad \text{and} \tag{3.7}$$

$$f_{\Lambda_k}(x_{\Lambda_k} | x_{\Lambda_k^c}) \approx e^{-|\Lambda_k| h(P)}. \tag{3.8}$$

where $h(P)$ is the specific entropy of P, and is positive except for the trivial cases. For the consistency of T_n, we need to have enough "empirical counts" $|\Lambda_n| R_{n,X}(A_k)$ for all sub-configurations x_{Λ_k}. Hence the expectation of $|\Lambda_n| R_{n,X}(A_k)$ under P, which is just $|\Lambda_n| P(x(k))$, needs to be large. Combining with (3.7), this means k should be the order of $\sqrt{\log n}$ at most. Meanwhile, the orders slower than $\sqrt{\log n}$ are not desirable because $\|f^{(k)} - f\|$ would tend to zero too slowly.

3.1.2 How to determine c?

Uniform upper and lower bounds for $f_{\Lambda_k}(x_{\Lambda_k} | x_{\Lambda_k^c})$ and $P(x(k))$, $x \in \Omega$ are derived by the following lemma.

Lemma 3.1. *There exist two constants* $0 < b_2 < b_1 < \infty$, *such that*

$$e^{-b_1 |\Lambda_k|} \leq P(x(k)) \leq e^{-b_2 |\Lambda_k|}, \quad \textit{and} \tag{3.9}$$

$$e^{-b_1 |\Lambda_k|} \leq f_{\Lambda_k}(x_{\Lambda_k} | x_{\Lambda_k^c}) \leq e^{-b_2 |\Lambda_k|}, \tag{3.10}$$

uniformly $\forall\, k \in \mathbb{N}$, $x \in \Omega$.

Proof. By the DLR equations [cf. Ruelle (1978)],

$$P(x(k)) = \int_{\Omega_{\Lambda_k^c}} f_{\Lambda_k}(x_{\Lambda_k} | y_{\Lambda_k^c})\, P(dy_{\Lambda_k^c}), \quad \forall\, x(k) \in \Omega_{\Lambda_k}. \tag{3.11}$$

Hence it suffices to show (3.10). For a bounded function g, let

$g^* = \sup g - \inf g$, then it follows from (2.4) that

$$(3.12)\quad f(x) = \{1 + \Sigma_{z \neq x_\omega} \exp[h(V(z) - V(x_\omega)) + \Sigma_{\xi \neq \omega} J(\xi)(U(z,x_\xi) - U(x_\omega,x_\xi))]\}^{-1}.$$

Therefore,

$$(3.13)\qquad e^{-b_1} \leq f(x) \leq e^{-b_2}, \quad \text{uniformly } \forall\, x \in \Omega,$$

where

$$(3.14)\qquad \begin{cases} b_1 = \log\{1 + (2r-1)\exp[|h|\, V^* + \Sigma_{\xi \neq \omega} |J(\xi)| U^*]\}\ ; \\ b_2 = \log\{1 + (2r-1)\exp[-|h| V^* - \Sigma_{\xi \neq \omega} |J(\xi)| U^*]\}\ . \end{cases}$$

Give an arbitrary order $1, \ldots, |\Lambda_k|$ to the set Λ_k and write $f_{\Lambda_k}(x_{\Lambda_k} | x_{\Lambda_k^c})$ as a product of conditional probabilities under P in an obvious way

$$f_{\Lambda_k}(x_{\Lambda_k} | x_{\Lambda_k^c}) = \prod_{i=1}^{|\Lambda_k|} P(X_i = x_i \mid X_j = x_j,\ j=1,\ldots,i-1;\ X_{\Lambda_k^c} = x_{\Lambda_k^c}).$$

By (3.13) and (2.7), we obtain (3.10) by taking further conditional expectation in each factor of $f_{\Lambda_k}(x_{\Lambda_k} | x_{\Lambda_k^c})$. □

Lemma 3.1 implies that for all $x(k) \in \Omega_{\Lambda_k}$, $|\Lambda_n| \cdot P(x(k))$ is bounded below by

$$|\Lambda_n| e^{-b_1 |\Lambda_k|} \approx n^2\, e^{-b_1\, c \log n} = n^{2 - b_1 c} \to \infty$$

as $n \to \infty$ provided $b_1 c < 2$.

In practice, since b_1 depends on the unknown parameters h and J, we need to make the assumption:

(A5) $\exists$ a known constant $b > 0$ such that $b_1 \leq b$, $\forall (h,J)$ of interest.

(A5) enables us to choose c such that

$$(3.15)\qquad 0 < bc < \frac{1}{5}\ .$$

Remark. (A1)-(A5) are sufficient conditions for the results in this paper, but not necessary. They merely simplify our argument and could be weakened. Among them, the knowledge of b in (A5) is particularly stringent, and to reduce it appears to be challenging. See Section 6 for further comment.

3.2. Consistency of T_n and $(\hat{h}_n;\ \hat{J}_n(\xi),\ \xi \in \mathbb{Z}^2)$.

The key in this paper is

Theorem 3.1. *Assume* (A1)-(A5) *and define* T_n *by* (3.5) *with* $k = [c \log n] \triangleq$ *the integer part of* $c \log n$, *where* c *satisfies* (3.15). *Then for every* $\epsilon > 0$, $\exists\ n_0 \in \mathbb{N}$, $C > 0$, $\alpha > 0$, $\beta > 0$, *such that*

$$P(\|T_n - f\| > \epsilon) \le C\, e^{-\alpha n^{\beta}}, \quad \forall\ n \ge n_0. \tag{3.16}$$

Here $P \in \mathcal{G}$ *is a GRF corresponding to* (h,J).

Remark. (3.16) gives the rate of consistency for T_n in (2.8). The proof will be given in Section 4.

Now we construct the estimator $(\hat{h}_n;\ \hat{J}_n(\xi),\ \xi \in \mathbb{Z}^2)$ via the method of pseudo-likelihood.

To illustrate the idea, suppose the pair-potential J has a finite range. Then $\theta \triangleq (h,J)$ is a finite-dimensional parameter, and the GRF P corresponding to θ is a MRF. For every $x \in \Omega$, define *the pseudo-likelihood function*

$$\mathcal{PL}_n(x;\theta) = \prod_{\xi \in \Lambda_n} f_\xi(x_\xi | {}_\xi x). \tag{3.17}$$

$\mathcal{PL}_n(x;\theta)$ is a concave function in θ. For the given sample X(n) and an arbitrary $x_{\Lambda_n^c}$, any $\hat{\theta}$ which maximizes $\mathcal{PL}_n(X(n) \oplus x_{\Lambda_n^c};\ \theta)$ is called a MPLE of θ. It is a good alternative of the MLE of θ, and has a great computational advantage over the MLE. In the model specified in Section 2, J is infinite-dimensional. The above method of pseudo-likelihood needs to be modified.

For the k chosen in Theorem 3.1, trancate (h,J) by

$$\theta_k \triangleq (h;\ J(\xi),\ \|\xi\| \le d_n = [\tfrac{k}{2}]). \tag{3.18}$$

Then similar to (2.1)-(2.4), define

$$H_{\{\xi\}}(x) = -h\ V(x_\xi) - \sum_{0<\|\xi-\eta\|\le d_n} J(\xi-\eta)\ U(x_\xi, x_\eta). \tag{3.19}$$

$$g_\xi(x;\theta_k) \triangleq g_\xi(x_\xi | x_\eta,\ 0 < \|\xi-\eta\| \le d_n;\ \theta_k) = Z^{-1}_{\{\xi\},\,\xi^x} \cdot e^{-H_\xi(x)}. \tag{3.20}$$

where $$Z_{\{\xi\},\,\xi^x} = \sum_{y_\xi} e^{-H_\xi(y_\xi \oplus_\xi x)}.$$

Note that $g_\xi(x;\theta_k)$ are the local characteristics that induce some MRF P' so that if P' generates X then $\forall\ x \in \Omega,\ \xi \in \mathbb{Z}^2$,

$$P'(X_\xi = x_\xi |_\xi X = {}_\xi x) = g_\xi(x;\theta_k). \tag{3.21}$$

In general, $P' \neq P$ and the local characteristics $f_\xi(x_\xi |_\xi x)$ may still depend on the entire sequence J. Nevertheless, for large n the two functions $f_\xi(\cdot)$, and $g_\xi(\cdot;\theta_k)$ are close to each other. This motivates the following construction.

Given the periodic configuration X^n, let

$$PL_n(X^n;\ \theta_k) = \prod_{\xi\in\Lambda_n} g_\xi(X^n;\theta_k),\quad \theta_k \in \Theta_K, \tag{3.22}$$

where Θ_K is a bounded, simple-connected open region in $\mathbb{R}^K$, with the closure $\bar{\Theta}_K$.

(3.23) $\hat{\theta}_k \triangleq (\hat{h}_n;\ \hat{J}_n(\xi),\ \|\xi\| \le d_n)$ be a maximizer of $PL_n(X^n,\ \theta)$ for $\theta \in \bar{\Theta}_K$, and

$$\hat{J}_n(\xi) = 0,\quad \forall\ \xi \in \mathbb{Z}^2 \text{ with } \|\xi\| > d_n. \tag{3.24}$$

Note that both θ_k and Θ_K depend on n. Apparently, for fixed n $PL_n(X^n;\theta)$ is a bounded continuous function in θ. Hence $\hat{\theta}_k$ exists on $\bar{\Theta}_K$, but need not be unique.

Theorem 3.2. *Assume* (A1)-(A5) *and let* $(\hat{h}_n;\ \hat{J}_n(\xi),\ \xi \in \mathbb{Z}^2)$ *be defined by* (3.23) *and* (3.24). *Then for every* $\epsilon > 0$, $\exists\ n' \in \mathbb{N}$, $C'' > 0$, $\mu > 0$, $\nu > 0$, *such that*

$$(3.25) \qquad P(|\hat{h}_n - h| + \sum_{\xi \in \mathbb{Z}^2} |\hat{J}_n(\xi) - J(\xi)| > \epsilon) \le C'' e^{-\mu n^{\nu}},\ \forall\ n \ge n'.$$

Again, $P \in \mathcal{G}$ *is induced by* (h,J).

Having Theorem 3.2 established, we obtain a solution to the problem (III) as well:

Corollary 3.1. *Suppose the sample* $X(n)$ *is generated by a MRF* P *induced by* (h,J) *with* $J(\xi) = 0$, $\forall\ \xi \in \mathbb{Z}^2$ *with* $\|\xi\| > d_n$.

Then

$$(3.26) \qquad f_\xi(x_\xi | {}_\xi x) = g_\xi(x;\theta_k),\ \forall\ \xi \in \mathbb{Z}^2,\ x \in \Omega.$$

And $(\hat{h}_n;\ \hat{J}_n(\xi),\ \xi \in \mathbb{Z}^2)$ *in* (3.23) *and* (3.24) *is a strongly consistent estimator of* (h,J). *Moreover,* (3.25) *holds.*

The proof of Theorem 3.2 will be given in Section 5.

4. The rate of consistency of T_n.

The aim of this section is to prove Theorem 3.1 which provides the rate of consistency for T_n. As we mentioned before, X may bear long-range dependence under $P \in \mathcal{G}$ due to the possible phase transitions. This difficulty can be overcome by studying the asymptotics of the conditional probability of some sub-configurations given their complements.

The following proposition may be called a "*conditional mixing lemma*".

Proposition 4.1. *Let* $B_1, \ldots, B_L$ *are bounded, simple-connected regions in* $\mathbb{Z}^2$, $\mathcal{B} = \mathbb{Z}^2 \setminus (\cup_{\ell=1}^{L} B_\ell)$. *Suppose the distance* $d(B_\ell, B_{\ell'}) \ge r_n$, $\forall\ \ell \ne \ell'$, *where* $\lim_{n\to\infty} r_n = \infty$ *and*

$$\lim_{n\to\infty} \max_{1\le \ell\le L} |B_\ell| \sum_{\|\xi\|\ge \gamma_n} |J(\xi)| = 0. \tag{4.1}$$

Then for any bounded measurable functions $u_\ell : \Omega_{B_\ell} \to \mathbb{R}$, $\ell = 1,\dots,L$, *we have*

$$E\left\{ \prod_{\ell=1}^{L} u_\ell(X_{B_\ell}) \,|x_{\mathcal{B}}\right\} = \left\{ \prod_{\ell=1}^{L} E[u_\ell(X_{B_\ell}) \,|x_{\mathcal{B}}]\right\} (1+\delta_n)^L \tag{4.2}$$

uniformly $\forall\, x_{\mathcal{B}} \in \Omega_{\mathcal{B}}$, *as* $n \to \infty$, *where* $E(\cdot\,|x_{\mathcal{B}})$ *is the conditional expectation with respect to* $P(\cdot\,|x_{\mathcal{B}})$, *and*

$$\delta_n = O\left[\max_{1\le \ell\le L} |B_\ell| \sum_{\|\xi\|\ge \gamma_n} |J(\xi)| \right], \quad \text{as} \quad n \to \infty. \tag{4.3}$$

Proof. Suppose K_1, K_2, K_3 form a disjoint decomposition of $\mathbb{Z}^2$ such that K_1 is bounded and simple-connected, $d(K_1,K_2) \ge \gamma_n$, and $K_3 = \mathbb{Z}^2\setminus(K_1 \cup K_2)$, then for all $x \in \Omega$, $x' \in \Omega$,

$$\begin{aligned} &|-H_{K_1}(x_{K_1} \oplus x_{K_3} \oplus x'_{K_2}) + H_{K_1}(x_{k_1} \oplus x_{K_3} \oplus x_{K_2})| \\ &\quad = \Big| \sum_{\zeta\in K_1} \sum_{\eta\in K_2} J(\zeta-\eta)[U(x_\zeta, x'_\eta) - U(x_\zeta, x_\eta)] \Big| \\ &\quad \le U^* |K_1| \sum_{\|\xi\|\ge \gamma_n} |J(\xi)|, \end{aligned}$$

where the notation $x_{K_1} \oplus x_{K_3} \oplus x_{K_2}$ means the configuration consisting of several parts, similar to that in (2.2). Hence

$$\begin{aligned} \frac{f_{K_1}(x_{K_1}|x_{K_3} \oplus x'_{K_2})}{f_{K_1}(x_{K_1}|x_{K_3} \oplus x_{K_2})} &= \exp\{-H_{K_1}(x_{k_1} \oplus x_{K_3} \oplus x'_{K_2}) + H_{K_1}(x_{K_1} \oplus x_{K_3} \oplus x_{K_2})\} \\ &\cdot \Sigma_{y_{K_1}} f_{K_1}(y_{K_1}|x_{K_3} \oplus x'_{K_2}) \cdot \exp\{-H_{K_1}(y_{K_1} \oplus x_{K_3} \oplus x_{K_2}) + H_{K_1}(y_{K_1} \oplus x_{K_3} \oplus x'_{K_2})\} \\ &= 1 + \Delta_n, \end{aligned} \tag{4.4}$$

where $|\Delta_n| \leq 2\, U^* |K_1| \sum_{\|\xi\| \geq \tau_n} |J(\xi)|$.

So $\forall\ \ell = 1,\ldots,L$,

$$(4.5) \qquad P(X_{B_\ell} = x_{B_\ell} | x_{\mathcal{B}} \oplus x_{B_1} \oplus \ldots \oplus x_{B_{\ell-1}}) = P(X_{B_\ell} = x_{B_\ell} | x_{\mathcal{B}})\ (1 + \delta_n),$$

as $n \to \infty$,

which implies

$$(4.6) \qquad P(X_{B_\ell} = x_{B_\ell},\ \ell = 1,\ldots,L | x_{\mathcal{B}}) = \left[\prod_{\ell=1}^{L} P(X_{B_\ell} = x_{B_\ell} | x_{\mathcal{B}}) \right] (1 + \delta_n)^L.$$

Therefore, (4.2) follows trivially. □

To create an environment for the application of the conditional mixing lemma, we decompose Λ_n in a "self-similar" fashion:

For technical convenience, we suppose $n = m^2$ for some $m \in \mathbb{N}$. Λ_n then is partitioned as m×m squares $D_1,\ldots,D_n$. Each D_i contains n sites, ordered by $1,\ldots,n$. All D_i's keep the same way of ordering. Hence every $\xi \in \Lambda_n$ is indexed by a pair (i,j), referred to as the j-th site in D_i, $i, j = 1,\ldots,n$. From (3.2)-(3.5), define

$$(4.7) \qquad \begin{cases} Y_{ij} = \delta_{t(i,j)_X n}(A_k)\ , \\ Y'_{ij} = \delta_{t(i,j)_X n}(A'_k), \quad i, j = 1,\ldots,n, \text{ and} \end{cases}$$

$$(4.8) \qquad \begin{cases} N_j = \sum_{i=1}^{n} Y_{ij}\ , \\ N'_j = \sum_{i=1}^{n} Y'_{ij}, \quad j = 1,\ldots,n. \end{cases}$$

Note that all Y_{ij}, Y'_{ij}, N_j, N'_j depend on n and x_{Λ_k}, and

(4.9) $$T_n(x) = \frac{N_1+\ldots+N_n}{N_1'+\ldots+N_n'} \quad \text{when} \quad \sum_{j=1}^{n} N_j > 0.$$

Let $N = \Sigma_{j=1}^{n} N_j$, $N' = \Sigma_{j=1}^{n} N_j'$, then

(4.10) $$I_{\{N>0\}} \left| \frac{N}{N'} - f^{(k)}(x) \right| \le \sum_{j=1}^{n} I_{\{N>0\}} \left| \frac{N_j}{N'} - \frac{N_j'}{N'} f^{(k)}(x) \right|$$

$$\le \sum_{j=1}^{n} I_{\{N_j>0\}} \frac{N_j'}{N'} \left| \frac{N_j}{N_j'} - f^{(k)}(x) \right| + \sum_{j=1}^{n} I_{\{N_j=0, N>0\}} \frac{N_j'}{N'} f^{(k)}(x).$$

Therefore,

(4.11) $$|T_n(x) - f(x)| \le \sum_{\ell=1}^{4} D_n^{(\ell)}(x), \quad \text{where}$$

$$D_n^{(1)}(x) = |f^{(k)}(x) - f(x)|;$$

$$D_n^{(2)}(x) = I_{\{N=0\}} \left|\tilde{a} - f^{(k)}(x)\right| ;$$

$$D_n^{(3)}(x) = \sum_{j=1}^{n} I_{\{N_j=0,\ N>0\}} \frac{N_j'}{N'} f^{(k)}(x) ;$$

$$D_n^{(4)}(x) = \sum_{j=1}^{n} I_{\{N_j>0\}} \frac{N_j'}{N'} \left| \frac{N_j}{N_j'} - f^{(k)}(x) \right| .$$

First of all, by setting $K_1 = \{\omega\}$ and $\tau_n = [\frac{k}{2}]$ in (4.4) we average out X'_{K_2} and obtain

(4.12) $$f^{(k)}(x) = f(x)(1 + \Delta_n), \quad \text{uniformly} \quad \forall\, x \in \Omega,$$

with $$|\Delta_n| \le 2\, U^* \cdot \sum_{\|\xi\| \ge [\frac{k}{2}]} |J(\xi)| \le C_0 \sum_{\|\xi\| \ge [\frac{k}{2}]} \|\xi\|\, e^{-a\|\xi\|} \to 0,$$

as $n \to \infty$ for some $C_0 > 0$. Hence

(4.13) $$\|D_n^{(1)}\| \to 0 \quad \text{as} \quad n \to \infty.$$

The next lemma is needed for $D_n^{(2)}(x)$, $D_n^{(3)}(x)$ and $D_n^{(4)}(x)$.

Lemma 4.1. *Let* $\lambda_n = n^{1-bc}$. *Then for every* $\epsilon \in (0,1)$, $\exists\, n_1 \in \mathbb{N}$, *such that*

$$P\left[\frac{N_j}{\lambda_n} < 1-\epsilon \right] \leq C_1 e^{-\alpha_1 n^{\beta_1}}, \quad \text{for some} \tag{4.14}$$

$C_1 > 0$, $\alpha_1 > 0$, $\beta_1 > 0$; *and uniformly* $\forall\, n \geq n_1$, $\forall\, j = 1,\ldots,n$, *and* $\forall\, x_{\Lambda_k}$.

Proof. Denote the kxk square lattice centered at (i,j) by $Q_{i,j}$, $i = 1,\ldots,n$. And let

$$Q_j = \mathbb{Z}^2 \setminus (\cup_{i=1}^n Q_{i,j}) \ . \tag{4.15}$$

Notice that Y_{ij} depends only on the sub-configuration on $Q_{i,j}$. And $\forall\, i_1 \neq i_2$,

$$d(Q_{i_1,j}, Q_{i_2,j}) \geq m-k. \tag{4.16}$$

For every x_{Q_j}, it follows from (4.2) that

$$E\left[e^{-N_j/\sqrt{\lambda_n}} \mid x_{Q_j} \right] = \left\{ \prod_{i=1}^n E\left[e^{-Y_{ij}/\sqrt{\lambda_n}} \mid x_{Q_j} \right] \right\} (1+\delta_n)^n, \tag{4.17}$$

with $\delta_n = O(k^2 \sum_{\|\xi\| \geq m-k} |J(\xi)|)$.

Furthermore, for large n by Taylor expansion

$$E\left[e^{-Y_{ij}/\sqrt{\lambda_n}} \mid x_{Q_j} \right] = 1 - \frac{1}{\sqrt{\lambda_n}} E(Y_{ij}|x_{Q_j}) + \frac{\psi}{2\lambda_n} E(Y_{ij}|x_{Q_j}) \tag{4.18}$$

$$(|\psi| \leq 1 \ \text{ depends on } Y_{ij})$$

$$\leq \exp\left\{ -\frac{1}{\sqrt{\lambda_n}} E(Y_{ij}|x_{Q_j}) \left(1 - \frac{\psi}{2\sqrt{\lambda_n}}\right) \right\}.$$

Since $E(Y_{ij}|x_{Q_j}) \geq e^{-b_1|\Lambda_k|} \geq n^{-bc}$ by (2.7), (3.10), (4.4) and (A5), we

derive from (4.17) that

$$P\left[\frac{N_j}{\lambda_n} < 1 - \epsilon \right] \le e^{(1-\epsilon)\sqrt{\lambda_n}} \cdot E\, e^{-N_j/\sqrt{\lambda_n}}$$

$$\le e^{(1-\epsilon)\sqrt{\lambda_n}} \cdot e^{-(1-\frac{\epsilon}{4})\sqrt{\lambda_n}} \cdot (1+\delta_n)^n$$

$$\le C_1\, e^{-\alpha_1 n^{\beta_1}}.$$

The uniformity with respect to j and x_{Λ_k} is obvious. □

Lemma 4.2. *For every* $\epsilon \in (0,1)$, $\exists\, n_2 \in \mathbb{N}$, *such that*

$$P(\|D_n^{(\ell)}\| > \tfrac{\epsilon}{4}) \le C_2\, e^{-\alpha_2 n^{\beta_2}}, \quad \ell = 2,3, \quad \text{for some} \tag{4.19}$$

$C_2 > 0$, $\alpha_2 > 0$, $\beta_2 > 0$, *and* $\forall\, n > n_2$.

Proof. By Lemma 4.1,

$$P(N_j = 0) \le P\left[\frac{N_j}{\lambda_n} < 1-\epsilon \right] \le C_1\, e^{-\alpha_1 n^{\beta_1}}, \quad \forall\, j = 1,\ldots,n.$$

Therefore,

$$P(\|D_n^{(2)}\| > \tfrac{\epsilon}{4}) \le (2r)^{|\Lambda_k|} P(N=0) \le (2r)^{c \log n}\, C_1\, e^{-\alpha_1 n^{\beta_1}}, \text{ and}$$

$$P(\|D_n^{(3)}\| > \tfrac{\epsilon}{4}) \le (2r)^{|\Lambda_k|} \sum_{j=1}^{n} P(N_j=0) \le (2r)^{c \log n} \cdot n \cdot C_1\, e^{-\alpha_1 n^{\beta_1}}.$$

Lemma 4.2 follows trivially. □

The similar conditioning argument can also apply to $D_n^{(4)}(\cdot)$ as shown in the following lemma.

Lemma 4.3. *For every* $\epsilon \in (0,1)$, $\exists\, n_3 \in \mathbb{N}$, *such that*

$$P(\|D_n^{(4)}\| > \tfrac{\epsilon}{4}) \le C_3\, e^{-\alpha_3 n^{\beta_3}}, \quad \text{for some} \tag{4.20}$$

$C_3 > 0$, $\alpha_3 > 0$, $\beta_3 > 0$, *and* $\forall\, n \ge n_3$.

Proof.

$$P\left[N_j > 0;\ \left|\frac{N_j}{N_j'} - f^{(k)}(x)\right| > \frac{\epsilon}{4}\right]$$

$$= P\left[N_j > 0;\ \left|\sum_{i=1}^{n} (Y_{ij} - f^{(k)}(x) \cdot Y_{ij}')\right| > \frac{\epsilon}{4} N_j'\right]$$

$$\le P\left[\frac{N_j}{\lambda_n} < 1-\epsilon\right] + P\left[\left|\sum_{i=1}^{n} W_{ij}\right| > \tau\lambda_n\right],$$

where $\tau = \frac{1}{4}\epsilon(1-\epsilon)$, $W_{ij} = Y_{ij} - f^{(k)}(x)\, Y_{ij}'$.

By (4.14), we only need to study the second term.

First by (4.4),

$$E(Y_{ij}|x_{Q_j}) = f^{(k)}(x)\, E(Y_{ij}'|x_{Q_j})\ (1 + \rho_n), \tag{4.21}$$

$$\text{where } \rho_n = O\left[\sum_{\|\xi\| \ge m-k} |J(\xi)|\right] = O\left[\sqrt{n}\, e^{-\frac{a}{2}\sqrt{n}}\right],$$

uniformly $\forall\, x \in \Omega$, $\forall\, j=1,\ldots,n$.

Taylor expansion implies

$$E\left[e^{W_{ij}|\sqrt{\lambda_n}} \mid x_{Q_j}\right] = 1 + \frac{1}{\sqrt{\lambda_n}} \cdot O(\rho_n) + \frac{\psi}{2\lambda_n} E(W_{ij}^2|x_{Q_j}) \tag{4.22}$$

$(|\psi| \le 1$ depends on $W_{ij})$.

Therefore,

$$P\left[\sum_{i=1}^{n} W_{ij} > \tau\lambda_n\right] = P\left[\frac{1}{\sqrt{\lambda_n}} \sum_{i=1}^{n} W_{ij} > \tau\sqrt{\lambda_n}\right] \tag{4.23}$$

$$\leq e^{-\tau\sqrt{\lambda_n}} \cdot E\left\{ \exp\left[\frac{1}{\sqrt{\lambda_n}} \sum_{i=1}^{n} W_{ij} \right] \right\}$$

$$\leq e^{-\tau\sqrt{\lambda_n}} \cdot E\left\{ \prod_{i=1}^{n} E\left[e^{W_{ij}/\sqrt{\lambda_n}} \mid X_{Q_j} \right] \right\} (1 + \delta_n)^n$$

$$\leq 2e^{-\tau\sqrt{\lambda_n}} \exp\{n \cdot \frac{1}{\sqrt{\lambda_n}} \cdot O(\rho_n) + \frac{n}{2\lambda_n} \}$$

(since $|\psi| \, E(W_{ij}^2 | x_{Q_j}) \leq 1$)

$$\leq 4 \exp\left[-\tau \, n^{\frac{1-bc}{2}} + \frac{1}{2} n^{bc} \right]$$

$$\leq 4 \, e^{-\alpha_4 n^{\beta_4}}, \quad \text{by (3.15).}$$

By the same token,

$$P\left[-\sum_{i=1}^{n} W_{ij} > \tau \lambda_n \right] \leq 4 \, e^{-\alpha_4 n^{\beta_4}} . \tag{4.24}$$

Finally, with (4.14), (4.23), (4.24) together we obtain that

$$P(\|D_n^{(4)}\| > \tfrac{\epsilon}{4}) \leq (2r)^{|\Lambda_k|} \cdot n \cdot C_5 \, e^{-\alpha_5 n^{\beta_5}} \leq C_3 \, e^{-\alpha_3 n^{\beta_3}} . \qquad \square$$

Thus we have completed *the proof of Theorem* 3.1 by combining (4.11), (4.13), (4.19) and (4.20).

The following strengthened result is needed in Section 5.

Corollary 4.1. *For every* $\epsilon' > 0$, *we have*

$$P\left[\sup_{x_{\Lambda_k}} |I_{\{N>0\}} \frac{N}{N'} - g(x;\theta_k)| > \frac{\epsilon'}{n^{bc}} \right] \leq C_6 \, e^{-\alpha_6 n^{\beta_6}} , \tag{4.25}$$

for some $C_6 > 0$, $\alpha_6 > 0$, $\beta_6 > 0$.

Proof. For every x_{Λ_k},

$$|I_{\{N>0\}} \frac{N}{N'} - g(x;\theta_k)| \le M_n^{(1)}(x) + M_n^{(2)}(x) + M_n^{(3)}(x), \text{ where}$$

$$M_n^{(1)}(x) = I_{\{N=0\}}\, g(x;\theta_k);$$

$$M_n^{(2)}(x) = \sum_{j=1}^{n} I_{\{N_j=0,\ N>0\}} \frac{N_j'}{N'} g(x;\theta_k);$$

$$M_n^{(3)}(x) = \sum_{j=1}^{n} I_{\{N_j>0\}} \frac{N_j'}{N'} \left| \frac{N_j}{N_j'} - g(x;\theta_k) \right| .$$

$M_n^{(2)}(x)$ and $M_n^{(3)}(x)$ are just $D_n^{(3)}(x)$ and $D_n^{(4)}(x)$ respectively with $f^{(k)}(x)$ replaced by $g(x;\theta_k)$.

If we also replace ϵ by $\frac{\epsilon'}{n^{bc}}$ in Lemma 1-3, then all those lemmas still hold. In particular, the constant τ in (4.23) and (4.24) becomes $O(n^{-bc})$. Hence (3.15) is needed to guarantee the exponential decay there. Since the argument would be exactly the same, we omit the details. □

5. The rate of consistency of $(\hat{h}_n;\ \hat{J}_n(\xi),\ \xi \in \mathbb{Z}^2)$.

In this section, we prove Theorem 3.2 by making use of the analysis carried out in Section 4 to refine the argument given in Geman and Graffigne (1986).

By (3.24) and (A2),

$$\sum_{\|\xi\|>d_n} |\hat{J}_n(\xi) - J(\xi)| = O(d_n\, e^{-ad_n}) \to 0 \quad \text{as} \quad n \to \infty. \tag{5.1}$$

Hence to obtain (3.25), it suffices to show that

$$P(\|\hat{\theta}_k - \theta_k\| > \epsilon) \le C''\, e^{-\mu n^{\nu}}, \tag{5.2}$$

where $\|\cdot\|$ is the ℓ^1-norm.

For $\theta \in \mathbb{R}^K$, define

$$G_n(\theta) = \sum_{x_{\Lambda_k \setminus \{\omega\}}} \frac{N'}{|\Lambda_n|} \sum_{x_\omega} g(x;\theta_k) \log \frac{g(x;\theta)}{g(x;\theta_k)} , \tag{5.3}$$

where $g(\cdot) = g_\omega(\cdot)$; and

$$F_n(\theta) = \sum_{x_{\Lambda_k \setminus \{\omega\}}} \frac{N'}{|\Lambda_n|} \sum_{x_\omega} \left[I_{\{N>0\}} \frac{N}{N'} \right] \log \frac{g(x;\theta)}{g(x;\theta_k)} . \tag{5.4}$$

Lemma 5.1 *Given the sample* $X(n)$ *and fixed* n,

(i) $G_n(\theta_k) = 0$; $G_n(\theta) \leq 0$, $\forall\ \theta$; $G_n(\theta)$ *is concave in* θ.

(ii) $F_n(\theta_k) = 0$; $F_n(\theta)$ *is concave in* θ; *and*

$$F_n(\theta) = \frac{1}{|\Lambda_n|} [\log PL_n(X^n;\theta) - \log PL_n(X^n;\theta_k)]. \tag{5.5}$$

Proof. $G_n(\theta_k) = 0$ and $F_n(\theta_k) = 0$ are obvious. $G_n(\theta) \leq 0$ follows from Jensen's inequality. The concavity of $G_n(\theta)$ and $F_n(\theta)$ follows from the concavity of $\log g(x;\theta)$ as a function of θ, which can be verified easily. Finally, by (4.7) and (4.8),

$$\sum_{x_{\Lambda_k \setminus \{\omega\}}} \frac{N'}{|\Lambda_n|} \sum_{x_\omega} \left[I_{\{N>0\}} \frac{N}{N'} \right] = \frac{1}{|\Lambda_n|} \sum_{(i,j)\in\Lambda_n} \sum_{x_{\Lambda_k}} \delta_{t(i,j)x^n}(\Lambda_k) = 1,$$

and $g_\xi(t^\xi x;\ \theta) = g(x;\theta)$, thus (5.5) holds. □

Now define the event

$$\mathcal{E}_n = \left\{ \frac{N'}{|\Lambda_n|} > \frac{1}{2n^{bc}} ,\ \forall\ x_{\Lambda_k} \right\}.$$

Lemma 5.2. $\exists\ q > 0$ *which depends only on* (h,J), V,U, *such that* $\forall$ *sufficiently large* n *and* $\forall\ \theta \in \mathbb{R}^K$,

$$\sup_{\|v\|=1} [v' \cdot \nabla^2 G_n(\theta) \cdot v] \leq -\frac{q}{n^{bc}} \quad \textit{and} \tag{5.6}$$

(5.7) $\sup_{\|v\|=1} [v' \cdot \nabla^2 F_n(\theta) \cdot v] \leq -\frac{q}{n^{bc}}$ *hold on* $\mathcal{E}_n$,

where $v \in \mathbb{R}^K$, v' *is the transpose of* v, *and* $\nabla^2 G_n(\theta)$, $\nabla^2 F_n(\theta)$ *are the Hessian matrices of* $G_n(\theta)$ *and* $F_n(\theta)$.

Proof. For notational convenience, let $s = x_{\Lambda_k \setminus \{\omega\}}$, and $\varphi(x_\omega, s) = (V(x_\omega); U(x_\omega, x_\xi), \|\xi\| \leq d_n)$. $\varphi(x_\omega, s)$ is formally regarded as a vector in $\mathbb{R}^K$. By standard calculation,

(5.8) $$v' \cdot \nabla^2 G_n(\theta) \cdot v = v' \cdot \nabla^2 F_n(\theta) \cdot u$$
$$= -\Sigma_s \frac{N'}{|\Lambda_n|} E_\theta\{(v'[\varphi(X_\omega, s) - E_\theta(\varphi(X_\omega, s)|s)])^2 | s\},$$

where $E_\theta(\cdot|s)$ is the conditional expectation on S with respect to $g_\omega(\cdot|s;\theta)$ in (3.20). Thus it suffices to show (5.6).

The vector $\varphi(x_\omega, s) - E_\theta[\varphi(X_\omega, s)|s]$ has the following components:

(5.9) $$V(x_\omega) - E_\theta[V(X_\omega)|s] = \sum_{z \in S} P(z|s)[V(x_\omega) - V(z)];$$

(5.10) $$U(x_\omega, x_\xi) - E_\theta[U(X_\omega, x_\xi)|s] = \sum_{z \in S} P(z|s)[U(x_\omega, x_\xi) - U(z, x_\xi)], \quad \xi \in \Lambda_k \setminus \{\omega\},$$

where $P(z|s) = P(X_\omega = z | X_{\Lambda_k \setminus \{\omega\}} = s) = g_\omega(z|s;\theta)$.

Given an arbitrary unit vector $v \in \mathbb{R}^K$, we can always find a sub-configuration x_{Λ_k} such that the corresponding components in v and in $\varphi(x_\omega, s) - E_\theta[\varphi(X_\omega, s)|s]$ have the same sign.

The scheme is as follows:

Assume (A3) and Case 2 in (A4). Let v_ω and v_ξ, $\xi \in \Lambda_k \setminus \{\omega\}$ be the components of v corresponding to (5.9) and (5.10) respectively. Set

$$\begin{cases} x_\omega = z_1,\ x_\xi = \bar{z}, & \text{if } v_\omega \geq 0,\ v_\xi \geq 0; \\ x_\omega = z_1,\ x_\xi = \underline{z}, & \text{if } v_\omega \geq 0,\ v_\xi < 0; \\ x_\omega = z_2,\ x_\xi = z^*, & \text{if } v_\omega < 0,\ v_\xi \geq 0; \\ x_\omega = z_2,\ x_\xi = z_*, & \text{if } v_\omega < 0,\ v_\xi < 0. \end{cases}$$

Case 1 of (A4) can be treated in the same way.

It follows from (A3), (3.13), (5.9) and (5.10) that

$$|v'(\varphi(x_\omega,s) - E_\theta[\varphi(X_\omega,s)|s])| \geq \|v\| e^{-b_1 q_1} > 0, \tag{5.11}$$

where $q_1 = \min\{V^*, \bar{U}(\cdot,\bar{z}), \bar{U}(\cdot,\underline{z}), \bar{U}(\cdot,z^*), \bar{U}(\cdot,z_*)\}$ with

$$\bar{U}(\cdot,\tilde{z}) \triangleq \max_{z\in S} U(z,\tilde{z}) - \min_{z\in S} U(z,\tilde{z}),\ \tilde{z} \in S.$$

Therefore, on the set $\mathcal{E}_n$, (5.6) follows from (5.8) and (5.11) by letting

$$q = \frac{1}{4} e^{-3b_1 q_1^2} > 0. \quad \square$$

Remark 5.1. Lemma 4.1 implies that

$$P(\mathcal{E}_n^c) \leq C_7 e^{-\alpha_7 n^{\beta_7}} \quad \text{for some } C_7 > 0,\ \alpha_7 > 0,\ \beta_7 > 0 \tag{5.12}$$

and $\forall$ sufficiently large n.

Therefore by Lemma 5.2, with an arbitrarily large probability both $G_n(\theta)$ and $F_n(\theta)$ are strictly concave in θ. In particular, θ_k would be the unique maximizer of $G_n(\theta)$.

Lemma 5.3. *For every $\epsilon > 0$ and $\forall$ sufficiently large n,*

$$\sup_{\|\theta-\theta_k\|=\epsilon} [G_n(\theta) - G_n(\theta_k)] < -\frac{\epsilon^2 q}{n^{bc}} \text{ holds on } \mathcal{E}_n. \tag{5.13}$$

Proof. By Taylor expansion,

$$\sup_{\|\theta-\theta_k\|=\epsilon} [G_n(\theta)-G_n(\theta_k)] = (\theta-\theta_k)' \nabla^2 G_n(\tilde{\theta})(\theta-\theta_k) \le \epsilon^2 \sup_{\|v\|=1} (v' \nabla^2 G_n(\tilde{\theta})v)$$

for some $\tilde{\theta}$ with $\|\tilde{\theta}-\theta_k\| \le \epsilon$. □

Lemma 5.4. *For every $\epsilon > 0$ and $\forall$ sufficiently large n,*

$$(5.14) \qquad P\left[\sup_{\|\theta-\theta_k\|=\epsilon} |F_n(\theta)-G_n(\theta)| > \frac{\epsilon^2 q}{n^{bc}} ; \mathcal{E}_n \right] \le C_8 e^{-\alpha_8 n^{\beta_8}},$$

for some $C_8 > 0$, $\alpha_8 > 0$, $\beta_8 > 0$.

Proof. Since $\left| \log \frac{g(x;\theta)}{g(x;\theta_k)} \right| \le 2b_1 \quad \forall\, \theta \in \mathbb{R}^K$,

$$\sup_{\|\theta-\theta_k\|=\epsilon} |F_n(\theta)-G_n(\theta)|$$

$$\le \sum_{x_{\Lambda_k \setminus \{\omega\}}} \frac{N'}{|\Lambda_n|} \sum_{x_\omega} \left| I_{\{N>0\}} \frac{N}{N'} - g(x;\theta_k) \right| \sup_{\|\theta-\theta_k\|=\epsilon} \left| \log \frac{g(x;\theta)}{g(x;\theta_k)} \right|$$

$$\le 4b_1 r \cdot \sup_{x_{\Lambda_k}} \left| I_{\{N>0\}} \frac{N}{N'} - g(x;\theta_k) \right|.$$

Hence (5.14) follows from Corollary 4.1 by letting $\epsilon' = \frac{\epsilon^2 q}{4b_1 r}$ in (4.25). □

Proof of Theorem 3.2. Take an arbitrary $\epsilon > 0$ such that $\{\theta : \|\theta-\theta_k\| \le \epsilon\} \subset \Theta_K$.

$$P(\|\hat{\theta}_k-\theta_k\| > \epsilon) \le P\left[\sup_{\|\theta-\theta_k\|=\epsilon} [F_n(\theta) - F_n(\theta_k)] > 0; \mathcal{E}_n \right] + P(\mathcal{E}_n^c)$$

$$\le P\left[\sup_{\|\theta-\theta_k\|=\epsilon} |F_n(\theta) - G_n(\theta)| > \frac{\epsilon^2 q}{n^{bc}} ; \mathcal{E}_n \right]$$

$$+ P\left[\sup_{\|\theta-\theta_k\|=\epsilon} [G_n(\theta) - G_n(\theta_k)] > -\frac{\epsilon^2 q}{n^{bc}} ; \mathcal{E}_n \right]$$

$$+ P(\mathcal{E}_n^c) .$$

(5.2) then follows from Lemma 5.4, Lemma 5.3 and (5.12). □

Remark 5.2. In general, the following *identifiability* condition needs to be imposed:

(A6) For every $n \in \mathbb{N}$, $\theta_k = \theta_k'$ holds whenever $g(x;\theta_k) = g(x;\theta_k')$ $\forall\, x \in \Omega$.

Nevertheless, (A6) is not used in our proof of Theorem 3.2 explicitly. (A6) is tightly connected with *the equivalence of potentials* [cf. Georgii (1988) and Gidas (1988)]. In fact, (A3), (A4) imply that each equivalence class of the pair-potentials is a singleton, hence (A6) holds.

In some special cases, consistent estimators for (h,J) can be computed directly without using MPLE.

Example 2.1. (*the GIM revisited*) Let

$$\pi_1 = (x_\zeta = 1, \quad \forall\, \zeta \in \Lambda_k);$$

$$\pi_2 = (x_\omega = 1;\ x_\zeta = -1, \ \forall\, \zeta \in \Lambda_k \backslash \{\omega\});$$

$$\pi_3 = (x_\zeta = -1, \ \forall\, \zeta \in \Lambda_k);$$

$$\pi_4 = (x_\omega = -1;\ x_\zeta = 1, \ \forall\, \zeta \in \Lambda_k \backslash \{\omega\});$$

$$\pi_{1\xi} = \pi_1, \quad \xi \in \Lambda_k \backslash \{\omega\};$$

$$\pi_{2\xi} = (x_\xi = -1;\ x_\zeta = 1, \ \forall\, \zeta \in \Lambda_k \backslash \{\xi\}), \quad \xi \in \Lambda_k \backslash \{\omega\};$$

$$\pi_{3\xi} = (x_\omega = -1, \ x_\xi = -1;\ x_\zeta = 1, \ \forall\, \zeta \in \Lambda_k \backslash \{\omega, \xi\}), \ \xi \in \Lambda_k \backslash \{\omega\};$$

$$\pi_{4\xi} = \pi_4, \quad \xi \in \Lambda_k \backslash \{\omega\}.$$

Then

$$h = \frac{1}{4} \log \frac{g(\pi_1;\theta_k) \cdot g(\pi_2;\theta_k)}{g(\pi_4;\theta_k) \cdot g(\pi_3;\theta_k)}\ ; \quad \text{and} \tag{5.15}$$

$$J(\xi) = \frac{1}{4} \log \frac{g(\pi_{1\xi};\theta_k) \cdot g(\pi_{3\xi};\theta_k)}{g(\pi_{4\xi};\theta_k) \cdot g(\pi_{2\xi};\theta_k)}\ ; \quad \xi \in \Lambda_k / \{\omega\}. \tag{5.16}$$

Based on these explicit expressions, the construction of $(\hat{h}_n;\ \hat{J}_n(\xi),\ \xi \in \mathbb{Z}^2)$ consists of two steps:

Step 1. Trancate (h,J) by θ_k as in (3.18) and estimate $g(x;\theta_k)$ by $T_n(x)$ defined in (3.5).

Step 2. Define

$$\hat{h}_n = \frac{1}{4} \log \frac{T_n(\pi_1) \cdot T_n(\pi_2)}{T_n(\pi_4) \cdot T_n(\pi_3)} ; \tag{5.17}$$

$$\hat{J}_n(\xi) = \begin{cases} \frac{1}{4} \log \dfrac{T_n(\pi_{1\xi}) \cdot T_n(\pi_{3\xi})}{T_n(\pi_{4\xi}) \cdot T_n(\pi_{2\xi})} , & \text{if } \xi \in \Lambda_k \backslash \{\omega\} ; \\ 0 , & \text{otherwise.} \end{cases} \tag{5.18}$$

Then the exponential rate of consistency for $(\hat{h}_n;\ \hat{J}_n(\xi),\ \xi \in \mathbb{Z}^2)$ in the sense of (3.25) can be derived by repeating the argument in Section 4.

6. Concluding Remarks.

This paper has provided a solution to the open problem of nonparametric estimation for GRF induced by pair-potentials. The results hold regardless of phase transition and symmetry breaking. The conditions (A1)-(A4) are satisfied in many examples of interest. They can be modified in many other image models without much difficulty, so that the argument in this paper still applies.

As mentioned in Section 1, one major unresolved issue is to replace (A5) by some data-driven procedures. Another important topic for future attention is to generalize the results to the models with more complicated degradation structure. It should also be pointed out that besides the grid size n and the range k of the neighboring system, there is a third factor in asymptotic analysis: the grey-level r. In practice r might also be large along with n and k. To explore the relation among n, k and r in various imaging problems should be very

interesting and challenging as well.

Acknowledgement. Grant support from ONR (N00014-89-J-1760) is greatfully acknowledged. The author particularly thanks Stuart Geman and Basilis Gidas for many stimulating discussions.

References

Besag, J. (1986). On the statistical analysis of dirty pictures (with discussion). *J. Roy. Stat. Soc., Series B*, **48**, 259-302.

Cométs, F. (1989). On consistency of a class of estimators for exponential families of Markov random fields on the lattice. Preprint, Univ. Paris-X.

Cométs, F. and Gidas, B. (1989). Parameter estimation for Gibbs distributions from partially observed data. Preprint, Brown Univ.

Föllmer, H. (1973). On entropy and information gain in random fields. *Z. Wahrsch. Verw. Geb.*, **26**, 207-217.

Geman, D. (1990). *Random Fields and Inverse Problems in Imaging*. To appear in *Lecture Notes in Math.*, Springer-Verlag, New York.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and Bayesian restoration of images. *IEEE Trans. PAMI-6*, 721-741.

Geman, S. and Graffigne, C. (1986). Markov random field image models and their applications to computer vision. *Proceedings of the International Congress of Mathematicians, 1986, ed. A.M. Gleason*, AMS, Providence.

Geman, S. and Hwang, C. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Stat.*, **10**, 401-414.

Geman, S. and McClure, D. (1987). Statistical methods for tomographic image reconstruction. *Proceedings of the 46th Session of the International Statistical Institute, Bulletin of the ISI*, **Vol. 52**.

Georgii, H.O. (1988). *Gibbs Measures and Phase Transitions*. Walter de Gruyter, Berlin-New York.

Gidas, B. (1986). Consistency of maximum likelihood and pseudo-likelihood estimators for Gibbs distributions. *Proceedings of the Workshop on Stochastic Differential Systems with Applications in Electrical/Computer Engineering, Control Theory, and Operations Research*, IMA, Univ. of Minnesota.

Gidas, B. (1988). Parameter estimation for Gibbs distributions. Preprint, Brown Univ.

Grenander, U. (1981). *Abstract Inference*. Wiley, New York.

Johnstone, I. and Silverman, B. (1990). Speed of estimation in positron emission tomography and related inverse problems. *Ann. Stat.*, 18, 251-280.

Ruelle, D. (1978). *Thermodynamic Formalism*. Addison-Wesley, Reading, Massachusetts.

Seymour, L. and Ji, C. (1990). Nearly optimal procedures for selecting Markov random fields in texture segmentation models. In preparation.