The Annals of Applied Probability
1992, Vol. 2, No. 1, 142–170

PARAMETER ESTIMATION FOR GIBBS DISTRIBUTIONS FROM PARTIALLY OBSERVED DATA

BY FRANCIS COMETS[1] AND BASILIS GIDAS[2]

Brown University

We study parameter estimation for Markov random fields (MRFs) over Z^d, $d \geq 1$, from incomplete (degraded) data. The MRFs are parameterized by points in a set $\Theta \subseteq \mathbb{R}^m$, $m \geq 1$. The interactions are translation invariant but not necessarily of finite range, and the single-pixel random variables take values in a compact space. The observed (degraded) process y takes values in a Polish space, and it is related to the unobserved MRF x via a conditional probability $P^{y|x}$. Under natural assumptions on $P^{y|x}$, we show that the ML estimations are strongly consistent irrespective of phase transitions, ergodicity or stationarity, provided that Θ is compact. The same result holds for noncompact Θ under an extra assumption on the pressure of the MRFs.

1. Introduction. The statistical inference for Gibbs distributions—equivalently, Markov random fields (MRFs)—has recently attracted a great deal of interest because of its importance in applications to image processing and computer vision tasks [17, 16, 5, 22, 19], neural modeling and perceptual inference [1, 25] and speech recognition [30, 3]. The inference problem has led [18, 23, 8] to an interesting interplay between statistics and the phenomena of phase transitions in statistical mechanics, and it generalizes the inference problem in time series analysis. Its fundamental difficulty lies in the presence of *long-range dependence* for the underlying random variables. In contrast to the situation in time series where short-range dependence is the rule and only special models are needed to exhibit long-range dependence, in MRF long-range dependence is typical, and it gives rise to the phenomena of phase transitions and the nonanalytic behavior of various thermodynamic quantities [36].

In some applications, the parameters of the Gibbs distributions need to be estimated from *fully observed data*, while in others from *incomplete* (*noisy*, *degraded*) *data*. Various methods have been devised for the case of fully observed data: (1) maximum likelihood (ML) estimation [15, 31, 38, 25]; (2) maximum pseudo-likelihood (MPL) estimation [4, 18, 21]; (3) the "coding" method [4]; (4) a logistic-like method [10, 35]; and (5) a "variational" method [2]. The main estimation procedures for the case of partially observed data are the ML method via the EM algorithm [9, 19] and the method of moments [19,

Received November 1989; revised January 1991.
[1]Research partially supported by AFOSR contract F-49620-86-C-0111.
[2]Research partially supported by ARO DAAL03-86-K-0110 and ONR contract N00014-88-K-0289.
AMS 1980 *subject classifications*. Primary 60F10; secondary 62F99.
Key words and phrases. Gibbs fields, large deviations, maximum likelihood, variational principle.

14]. A simple (EM-like) procedure for solving certain moment equations has recently been introduced in [2].

Consistency and the asymptotic behavior of estimators in the case of fully observed data have recently been studied in detail: Geman and Graffigne [18] provided the first proof of consistency for MPL estimators (see [21] for an alternative proof; see also [24]). In [23], it was shown that ML estimators are (strongly) consistent irrespective of phrase transitions, ergodicity or stationarity (some consistency results for the Ising model have been obtained in [33]). It was also shown in [23] that, under appropriate conditions, ML estimators are asymptotically normal and efficient. A statistical analysis of the Gaussian Markovian case is given in [27]. In [8], we established a superefficiency phenomenon for the Curie–Weiss model. A similar phenomenon is expected to hold for the Ising and other models.

In this paper we prove (strong) consistency of ML estimators for the case of incomplete (noisy) data. Our results hold irrespective of phase transitions, ergodicity or stationarity. The proof of consistency for incomplete data is much subtler than the proof [23] of consistency for fully observed data. An important step toward establishing consistency is the proof of a new variational principle for the conditional pressure (Section 3)—a result of independent interest. The proof of consistency also involves certain large deviations estimates [13, 32, 7] for the empirical field of the degraded data. After the completion of the present paper, we learned that a weaker consistency result, under stronger assumptions, and by difference methods, was obtained in [39, 40].

Our precise framework and consistency result are given in Section 2 (and proofs in Sections 3 and 4). Here we provide a brief outline only: The Gibbs distributions are parametrized by points in a parameter space Θ which is a subset of a finite-dimensional Euclidean space $\mathbb{R}^m$, $m \geq 1$. The interactions are translation invariant but not necessarily of finite range, and the single-pixel random variables ("spins") x_i, $i \in Z^d$, take values in a finite or compact state space $\Omega_{0,x}$. The state space for the MRFs over Z^d is $\Omega_x = (\Omega_{0,x})^{Z^d}$. The points (configurations) in Ω_x will be denoted by $x = \{x_i: i \in Z^d\}$. The process x is observed indirectly through an observable process $y = \{y_i: i \in Z^d\}$, where each y_i is assumed to take values in some Polish (i.e., complete separable metric) space $\Omega_{0,y}$. The state space of the observed process y is $\Omega_y = (\Omega_{0,y})^{Z^d}$. The unobserved and the observed processes are related through a known (independent of $\theta \in \Theta$) conditional probability $P^{y|x}$. Our general model (see Section 2) for $P^{y|x}$ covers degradations due to linear blurring, nonlinearities, noise, and so forth. For simplicity, we assume here that $P^{y|x}$ has the form

$$P^{y|x} = \left(\mu_0^{y_i|x_i}\right)^{\otimes Z^d}, \tag{1.1}$$

where $\mu_0^{y_i|x_i}$, $i \in Z^d$, is a (known) single-pixel conditional probability for y_i given x_i (this model covers, for example, the case when y_i is obtained from x_i by an additive or multiplicative noise η_i which is stochastically independent of x_i, e.g., $y_i = f(x_i) + \eta_i$, where f is a nonlinear transformation). If $\pi_\theta = \pi_\theta^x$ is a Gibbs distribution for the unobserved process, then $P^{y|x} \otimes \pi_\theta^x$ is the joint distribution of (x, y). The marginal of the observed process $y \in \Omega_y$ will be denoted by $P_\theta = P_\theta^y$.

We are interested in estimating the vector-parameter θ of the Gibbs distributions from a single observation $y(\Lambda) = \{y_i\colon i \in \Lambda\}$ in a finite window ("volume") $\Lambda \subset Z^d$, and then studying consistency as $\Lambda \to Z^d$. In Section 2, we will consider various log-likelihood functions. Here we consider a log-likelihood function based on "finite-volume" Gibbs distributions with "free boundary conditions" (see Section 2): Let $\mu_{0,x}$ be a probability measure on $\Omega_{0,x}$. The finite-volume Gibbs distribution with free boundary conditions in the finite window $\Lambda \subset Z^d$ has the form

$$\pi_{\Lambda,\theta}(dx(\Lambda)) = \frac{e^{\theta \cdot U_\Lambda(x(\Lambda))}}{Z_\Lambda(\theta)} \prod_{i\in\Lambda} \mu_{0,x}(dx_i), \tag{1.2}$$

where $x(\Lambda) = \{x_i\colon i \in \Lambda\}$, $U_\Lambda(x(\Lambda))$ are the energies (see Section 2) in the window Λ and $Z_\Lambda(\theta)$ is a normalizing constant, called the partition function. Under $\mu_0(dx_i, dy_i) = \mu(dy_i|x_i)\mu_{0,x}(dx_i)$, the marginal of y_i will be denoted by $\mu_{0,y}(dy_i)$, and the conditional probability of x_i given y_i will be denoted by $\mu_0(dx_i|y_i)$. The law of $y(\Lambda) = \{y_i\colon i \in \Lambda\}$ has a density $P_{\Lambda,\theta}(y(\Lambda))$ with respect to $\prod_{i\in\Lambda}\mu_{0,y}(dy_i)$, and the log-likelihood function is taken to be

$$\begin{aligned} l_n(y(\Lambda),\theta) &= -\frac{1}{|\Lambda|}\log P_{\Lambda,\theta}(y(\Lambda)) \\ &= p_\Lambda(\theta) - p_\Lambda(y(\Lambda),\theta), \end{aligned} \tag{1.3}$$

where $p_\Lambda(\theta) = (1/|\Lambda|)\log Z_\Lambda(\theta)$ is the finite-volume pressure and $p_\Lambda(y(\Lambda);\theta) = (1/|\Lambda|)\log Z_\Lambda(y(\Lambda),\theta)$ is the finite-volume "conditional" pressure [see (2.12)].

Both $p_\Lambda(\theta)$ and $p_\Lambda(y;\theta)$ are convex in θ, but $l_\Lambda(y,\theta)$ is not convex in θ. This is in contrast to the situation in the case of fully observed data [23]. Many of the difficulties and subtleties in the proof of consistency lie in the behavior of $p_\Lambda(y,\theta)$ as $\Lambda \to Z^d$. If θ_0 is the true parameter, and if the true distribution $P_{\theta_0}(dy)$ is translation invariant, then $p_\Lambda(y,\theta)$ has an a.s. limit $p(\cdot,\theta)$ as $\Lambda \to Z^d$ [Theorem 3.1(i)]. If $P_{\theta_0}(dy)$ is ergodic, then $p(\cdot,\theta)$ is a constant which satisfies a variational principle [Theorem 3.1(ii)]. This variational principle is a key step toward proving consistency; it is related to the asymptotics of Gibbs distributions under conditioning by P_{θ_0}, but the strategy used in [7] does not work here, because of the complex structure of P_{θ_0}. If $P_{\theta_0}(dy)$ is not translation invariant, then $p_\Lambda(y,\theta)$ need not converge as $\Lambda \to Z^d$. However, using large deviations estimates for the empirical field

$$R_{\Lambda,y} = \frac{1}{|\Lambda|}\sum_{i\in\Lambda}\delta_{\tau^i y}, \tag{1.4}$$

where τ^i is the shift on Z^d, we show that the limit points of $p_\Lambda(y,\theta)$ lie in a certain region determined by the ergodic distributions associated with θ_0. This result together with the above-mentioned variational principle are the major ingredients for establishing consistency when the parameter space Θ is compact (Theorem 2.1). When Θ is noncompact, the nonconvexity in θ of the log-likelihood function $l_\Lambda(y;\theta)$ creates subtle difficulties in showing that the minimizer of (1.3) exists for large Λ, and eventually stays in a compact subset of Θ. Our result of consistency in the noncompact case (Theorem 2.2) is the same as in the compact case, but it holds under a condition [assumption (2.21)] on the behavior of the pressure for large $|\theta|$. This condition is proven in the Appendix in a special case (which covers the Ising model without an external field), and it holds in general whenever $p_\Lambda(\theta)$ has an asymptote uniformly in the volume Λ. We believe that consistency in the noncompact case does not hold in complete generality without any extra assumption such as (2.21). In the Appendix, we argue that this problem and its difficulty are related to the fact that the set of *ground random fields* [20] (i.e., the set of MRFs with $|\theta| = +\infty$) is in general [20] larger than the set of the *attainable ground random fields* (i.e., the set of limit points of MRFs as $|\theta| \to +\infty$). We note that consistency for noncompact Θ (and incomplete data) does not seem to have been treated in the literature even for i.i.d. random variables (see [37] for a study of the i.i.d. case with compact Θ).

Our assumption that the single-pixel state space $\Omega_{0,x}$ (and hence Ω_x) is compact is not necessary. Our result holds for an arbitrary Polish space $\Omega_{0,x}$, provided that the summability condition (2.3) holds. However, this condition excludes the natural framework of unbounded spin systems [28]. Our techniques apply to noncompact $\Omega_{0,x}$, but they require certain technical modifications.

The organization of the paper is as follows: In Section 2, we set up our precise framework and state our consistency result. Section 3 contains some technical propositions, the variational principle for the conditional pressure and the proof of consistency for compact Θ. The proof of consistency for noncompact Θ is given in Section 4. Finally, the Appendix contains a proof of assumption (2.21) in a special case, and some miscellaneous remarks pertaining to the consistency for noncompact Θ.

2. Notation and main results. In this section, we set up our notation, summarize some properties of the Gibbs distributions and state our main result. Proofs are given in Section 3.

2.1. *Gibbs distributions.* We follow the notation of [36, 23]. Let $\Omega_{0,x}$ be the single-pixel state space, assumed to be a finite set or a compact space. With each pixel $i \in Z^d$, we associate a random variable ("spin") x_i taking values in $\Omega_{0,x}$. We set $\Omega_x = (\Omega_{0,x})^{Z^d}$, and $\Omega_{V,x} = (\Omega_{0,x})^V$ for any subset $V \subset Z^d$.

Gibbs distributions are defined in terms of *interactions*. An interaction Φ is a real continuous map

$$\Phi\colon \bigcup_{V \subset Z^d \text{ finite}} \Omega_{V,x} \to \mathbb{R},$$

and $\Phi(x(V))$ describes the interaction inside the subset V. Let Λ be a finite subset of Z^d, $\Lambda^c = Z^d \setminus \Lambda$, and z a fixed configuration in Ω_x. The *energy* in Λ

with *boundary conditions* (b.c.) z is $-U_{\Lambda,z}(x(\Lambda))$, where

$$U_{\Lambda,z}(x(\Lambda)) = \sum_{V\subset\Lambda} \Phi(x(V)) + \sum_{V\subset Z^d}{}' \Phi(x(V)\vee z(V)), \tag{2.1}$$

where the sum Σ' extends over finite $V \subset Z^d$ such that $V\cap\Lambda \neq \varnothing$, $V\cap\Lambda^c \neq \varnothing$, and the configuration $x(\Lambda)\vee z(V)$ is defined by

$$(x(V)\vee z(V))_i = \begin{cases} x_i, & \text{if } i\in\Lambda\cap V,\\ z_i, & \text{if } i\in\Lambda^c\cap V.\end{cases} \tag{2.2}$$

Free b.c. correspond to (2.1) without the second term.

In this paper the interactions are assumed to be translation invariant, that is, $\Phi(x(i+V)) = \Phi(x(V))$, and to satisfy

$$\|\Phi\| = \sum_{0\in V\subset Z^d \text{ finite}} \sup_{x(V)} |\Phi(x(V))| < +\infty. \tag{2.3}$$

If $\Phi(x(V)) = 0$ whenever the diameter of V is larger than R_0, then we say that Φ is a *finite-range* interaction of interaction radius R_0. The set of translation-invariant interactions that satisfy (2.3) form a separable Banach space $\mathscr{B}$. The set $\mathscr{B}_0$ of finite-range interactions is dense in $\mathscr{B}$ [36].

In this paper, we fix $m\geq 1$ interactions Φ^α, $\alpha = 1,\dots,m$, in $\mathscr{B}$ (with corresponding energy functions $U_{\Lambda,z}^{(\alpha)}$, $\alpha = 1,\dots,m$) and parametrize the Gibbs distributions by $\theta = (\theta^{(1)},\theta^{(2)},\dots,\theta^{(m)})\in\Theta$, where the parameter space is taken to be an arbitrary subset of $\mathbb{R}^m$. We will use the norm

$$\|U\|^2 = \sum_{\alpha=1}^{m} \|\Phi^{(\alpha)}\|^2. \tag{2.4}$$

Note that the energies can be written as follows:

$$U_{\Lambda,z}^{(\alpha)}(x(\Lambda)) = \sum_{i\in\Lambda}\ \sum_{i\in V\subset Z^d \text{ finite}} \frac{\Phi^\alpha(x(V)\vee z(V))}{|V\cap\Lambda|}.$$

Later we will use the functions ($x\in\Omega_x$),

$$A_{U^{(\alpha)}}(x) = \sum_{0\in V\subset Z^d \text{ finite}} \frac{\Phi^\alpha(x(V))}{|V|}, \tag{2.5a}$$

$$A_U = \{A_{U^\alpha}(x)\}_{\alpha=1}^{m}. \tag{2.5b}$$

Let $\mu_{0,x}$ be a probability measure on $\Omega_{0,x}$ and set

$$\mu_{0,x}^{(\Lambda)}(dx(\Lambda)) = \prod_{i\in\Lambda} \mu_{0,x}(dx_i). \tag{2.6}$$

The *finite-volume Gibbs distribution* in the finite window $\Lambda\subset Z^d$ with b.c. z is defined by

$$\pi_{\Lambda,\theta,z}(dx(\Lambda)) = \frac{e^{\theta\cdot U_{\Lambda,z}(x(\Lambda))}}{Z_{\Lambda,z}(\theta)}\mu_{0,x}^{(\Lambda)}(dx(\Lambda)). \tag{2.7}$$

An *infinite-volume Gibbs distribution*, or simply a *Gibbs distribution* associated with the (fixed) interactions $\Phi^{(\alpha)}$, $\alpha = 1, \dots, m$, and parametrized by $\theta \in \Theta$, is a probability measure π_θ on Ω_x whose conditional probability that $x|_{\Omega_{\Lambda,x}} = x(\Lambda)$ when it is known that $x|_{\Omega_{\Lambda^c,x}} = x(\Lambda^c)$ is given by

$$\pi_\theta(dx(\Lambda)|x(\Lambda^c)) = \pi_{\Lambda,\theta,x(\Lambda^c)}(x(\Lambda))\mu_{0,x}^{(\Lambda)}(dx(\Lambda)) \tag{2.8}$$

for every finite $\Lambda \subset Z^d$. The set of Gibbs distributions corresponding to the parameter-vector θ (and fixed Φ^α, $\alpha = 1, \dots, m$) will be denoted by $G(\theta)$. It is well known [36] that $G(\theta)$ is a convex, compact and Choquet simplex. If $G(\theta)$ is not a singleton, we say that a phase transition occurs for the parameter value θ. $G(\theta)$ always contains translation-invariant measures, but it may also contain [12] nontranslation-invariant distributions.

2.2. *The observed process.* The process $x = \{x_i: i \in Z^d\}$ is observed indirectly through an observable process $y = \{y_i: i \in Z^d\}$, where each y_i takes values in a Polish (i.e., complete separable metric) space $\Omega_{0,y}$. The state space for $y = \{y_i: i \in Z^d\}$ is $\Omega_y = (\Omega_{0,y})^{Z^d}$. The unobserved process x and the observed process y are related through a known (and independent of θ) conditional probability $P^{y|x}$. If $\pi_\theta \in G(\theta)$, then the joint distribution of (x, y) is $P^{y|x} \otimes \pi_\theta$. The marginal distribution of y will be denoted by P_θ^y or simply by P_θ. The set of Gibbs distributions $G(\theta)$ gives rise to a set $K(\theta)$ of probability distributions for the observed process y. Clearly, $K(\theta)$ is also convex and compact.

Throughout the paper we will use the notation $\Omega = \Omega_x \times \Omega_y$, $\Omega_0 = \Omega_{0,x} \times \Omega_{0,y}$, and for any subset $V \subset Z^d$, $\Omega_V = \Omega_0^V = \Omega_{V,x} \times \Omega_{V,y}$, $\Omega_{V,x} = \Omega_{0,x}^V$, $\Omega_{V,y} = \Omega_{0,y}^V$.

Most of our results go through by assuming only that $P^{y|x}$ is chosen so that the pair (x, y) is a *Markov random field*, although our variational principle [Theorem 3.1(ii)] does not need this property. However, we will assume, for simplicity, that $P^{y|x}$ has the following form: Let W be a fixed neighborhood of $0 \in Z^d$. Then

$$P^{y|x}(dy|x) = \prod_{i \in Z^d} P^{y|x}(dy_i|x(i + W)). \tag{2.9a}$$

We will also assume that $P^{y|x}(dy_i|x(i + W))$ has the following structure: Let $\mu_0(\cdot\,|\,\cdot)$ be a transition probability kernel from $\Omega_{0,x}$ to $\Omega_{0,y}$; the marginal of y_i under $\mu_0(dx_i, dy_i) \equiv \mu_0(dy_i|x_i)\mu_{0,x}(dx_i)$ will be denoted by $\mu_{0,y}(dy_i)$—a probability measure on $\Omega_{0,y}$. We will assume that $P^{y|x}(dy_i|x(i + W)) \ll \mu_0(dy_i|x_i)$ and

$$P^{y|x}(dy_i|x(i + W)) = e^{\Psi_0(x(i+W), y_i)}\mu_0(dy_i|x_i) \tag{2.9b}$$

with Ψ_0 some real, continuous, bounded map on $\Omega_{0,x}^W \times \Omega_{0,y}$. We now define a new interaction function $\Psi: \bigcup_{V \subset Z^d} \Omega_V \to \mathbb{R}$ by

$$\Psi(x(V), y(V)) = \begin{cases} \Psi_0(x(i + W), y_i), & \text{if } V = i + W \text{ for some } i \in Z^d, \\ 0, & \text{otherwise,} \end{cases}$$

and for finite $\Lambda \subset Z^d$,

(2.10a) $$\Psi_\Lambda(x(\Lambda), y(\Lambda)) = \sum_{V \subset \Lambda} \Psi(x(V), y(V)),$$

(2.10b) $$\Psi_{\Lambda,z}(x(\Lambda), y(\Lambda)) = \sum_{V \subset W(\Lambda)} \Psi(x(V), y(V)) = \sum_{i \in \Lambda} \Psi_0(x_{W(i)}, y_i),$$

where $W(\Lambda) + \Lambda + W$, and $x_{W(\Lambda)} = x(\Lambda) \vee z$. Then

(2.11) $$P^{y|x}(dy(\Lambda)|x) = \exp\{\Psi_{\Lambda, x(\Lambda^c)}(x(\Lambda), y(\Lambda))\} \prod_{i \in \Lambda} \mu_0(dy_i|x_i).$$

The marginal distribution of

$$y(\Lambda) = \{y_i : i \in \Lambda\} \quad \text{under} \quad P^{y|x}(dy(\Lambda)|x(\Lambda) \vee z) \cdot \pi_{\Lambda,\theta,z}(dx(\Lambda))$$

is given by

(2.12a) $$P_{\Lambda,\theta,z}(dy(\Lambda)) = P_{\Lambda,\theta,z}(y(\Lambda))\mu_{0,y}^{\Lambda}(dy(\Lambda)),$$

where $\mu_{0,y}^{\Lambda} = \mu_{0,y}^{\otimes\Lambda}$, and

(2.12b) $$P_{\Lambda,\theta,z}(y(\Lambda)) = \int \frac{1}{Z_{\Lambda,z}(\theta)} \exp\{\theta \cdot U_{\Lambda,z}(x(\Lambda)) + \Psi_{\Lambda,z}(x(\Lambda), y(\Lambda))\} \times \prod_{i \in \Lambda} \mu_0(dx_i|y_i)$$

(2.12c) $$= \frac{Z_{\Lambda,z}(y(\Lambda), \theta)}{Z_{\Lambda,z}(\theta)},$$

where $Z_{\Lambda,z}(y(\Lambda), \theta)$ is the conditional partition function given by

$$Z_{\Lambda,z}(y(\Lambda), \theta) = \int \exp\{\theta \cdot U_{\Lambda,z}(x) + \Psi_{\Lambda,z}(x,y)\} \prod_{i \in \Lambda} \mu_0(dx_i|y_i).$$

For the function Ψ we will have the analog of (2.3), that is,

(2.13) $$\|\Psi\| = \sum_{0 \in V \subset Z^d \text{ finite}} \sup_{x,y} |\Psi(x(V), y(V))| = \|\Psi_0\|_\infty < +\infty.$$

As in (2.5) we define

$$A_\Psi(x,y) = \sum_{0 \in V \subset Z^d \text{ finite}} \frac{\Psi(x(V), y(V))}{|\Lambda|} = \frac{1}{|W|} \sum_{i \in W} \Psi_0(x(i + W), y_i).$$

REMARKS.

1. One can easily verify that under the model (2.11), the pair (x, y) is a Markov random field with interaction $\theta \cdot \Phi + \Psi$. Thus, if Φ has finite range with interaction radius R_0, then $\theta \cdot \Phi + \Psi$ also has finite range with interaction radius $\max(R_0, \operatorname{diam} W)$, where $\operatorname{diam} W$ denotes the diameter of the neighborhood W.

2. Although x and (x, y) are MRFs, y is *not* an MRF.
3. Model (2.11) covers incomplete data situations and general degradation mechanisms such as blurring, nonlinear deformations and noise. In particular, it covers degradations of the form $y_i = f((Hx)_i, \eta_i)$, $i \in Z^d$, where H is a blurring matrix of spread W, η is a white noise (independent of x) and f is a nonlinear function such that the distribution of y_i given x is of the form $\exp\{\tilde{\Psi}_0(x(W), y_i)\}\, d\nu_0(y_i)$ with some continuous bounded $\tilde{\Psi}_0$ and some probability measure ν_0. The product structure in (2.11) does not cover the case of Markovian noise $\eta = \{\eta_i\colon i \in Z^d\}$, but our procedure can be modified to cover such cases.
4. Condition (2.13) is restrictive, but it can be relaxed. It does cover models of the form $y_i = f(x_i) \oplus \eta_i$ for a large class of additive white noise η_i (including Gaussian noise) and some multiplicative noise models, but it does not cover models of the form $y_i = (Hx)_i + \eta_i$ with η_i, say, in $\mathbb{R}$. Our techniques can be extended to treat such models by considering the setup of superstable interactions [27].
5. Our degradation model may be slightly modified to cover the case when the parameters of the noise η are unknown: Replace the process x in (2.8) by (x, η).

2.3. *Log-likelihood functions.* For each boundary condition z, we define a log-likelihood function in terms of (2.12), that is,

$$(2.14)\quad l_{\Lambda,z}(y(\Lambda);\theta) = -\frac{1}{|\Lambda|}\log P_{\Lambda,\theta,z}(y(\Lambda)) = p_{\Lambda,z}(\theta) - p_{\Lambda,z}(y(\Lambda),\theta),$$

where $p_{\Lambda,z}(\theta)$ and $p_{\Lambda,z}(y(\Lambda),\theta)$ are the pressure and the conditional pressure defined by

$$p_{\Lambda,z}(\theta) = \frac{1}{|\Lambda|}\log Z_{\Lambda,z}(\theta),$$

$$p_{\Lambda,z}(y(\Lambda),\theta) = \frac{1}{|\Lambda|}\log Z_{\Lambda,z}(y(\Lambda),\theta).$$

We define a second log-likelihood function as follows: For a distribution $P_\theta \in K(\theta)$, we denote by $P_\theta^{(\Lambda)}$ its restriction to $\Omega_{\Lambda,y}$, and by $f_\Lambda(y(\Lambda);\theta)$ the Radon–Nikodym derivative of $P_\theta^{(\Lambda)}$ with respect to $\mu_{0,y}^{(\Lambda)}(dy(\Lambda))$ (it is easily seen that f_Λ exists). The second log-likelihood function reads

$$(2.15)\qquad \tilde{l}_\Lambda(y(\Lambda);\theta) = \frac{1}{|\Lambda|}\log f_\Lambda(y(\Lambda);\theta).$$

From the computational point of view, (2.14) is more tractable than (2.15), but from the mathematical point of view (2.15) is a natural log-likelihood function. Our consistency theorems hold for both log-likelihood functions.

The sequence (net) of observations $y(\Lambda)$ in an expanding sequence (net) of windows $\Lambda \subset Z^d$ may arise in two ways: (1) There is an underlying infinite sample $y = \{y_i\colon i \in Z^d\}$, and we observe larger and larger pieces $y(\Lambda) =$

$\{y_i: i \in \Lambda\}$ of it. (2) The net $\{y(\Lambda)\}$ is a net of samples, possibly independent, from the net $\{P_{\Lambda,\theta,z}\}$ (with the same θ, but not necessarily the same boundary conditions). In the former case we will write, for example, $l_{\Lambda,z}(y,\theta)$ instead of $l_{\Lambda,z}(y(\Lambda),\theta)$.

2.4. *Main results.* In proving consistency of ML estimators, we will assume *identifiability* in the following sense:

(2.16) We say that $\theta_0 \in \Theta$ is *identifiable* if $\theta \neq \theta_0$ implies $K(\theta) \cap K(\theta_0) = \varnothing$.

Condition (2.16) clearly implies that if $\theta \neq \theta_0$, then $G(\theta) \cap G(\theta_0) = \varnothing$ (which is the identifiability condition for fully observed data [23]). The converse is not true. For example, for the standard binary Ising model without external field, if the observed data y are such that $y_i = x_i \eta_i$, where η_i is a fair Bernoulli process, that is, $P(\eta_i) = \frac{1}{2}\delta_{\eta_i - 1} + \frac{1}{2}\delta_{\eta_i + 1}$, then the distribution of $y(\Lambda) = \{y_i: i \in \Lambda\}$ is $P(y(\Lambda)) = 1/2^{|\Lambda|}$, independent of the parameters of the Ising model. Clearly, in this case the parameter of the Ising model (i.e., the temperature) cannot be estimated from the observed data $y = \{y_i: i \in Z^d\}$.

The following theorem is our consistency theorem for the case of compact parameter space $\Theta \in \mathbb{R}^m$.

THEOREM 2.1. *Let $\theta_0 \in \Theta$ be the true parameter vector, and let P_{θ_0} be any distribution in $K(\theta_0)$. Assume that Θ is compact and θ_0 identifiable. Let $\hat{\theta}_{\Lambda,z}$ be a measurable minimizer of $l_{\Lambda,z}(y(\Lambda),\theta)$. Then independently of the b.c. z, we have*

$$\hat{\theta}_{\Lambda,z} \to \theta_0, \quad P_{\theta_0}\text{-a.s., as } \Lambda \to Z^d.$$

Furthermore, for all $\varepsilon > 0$, we have

$$P_{\theta_0}\{|\hat{\theta}_{\Lambda,z} - \theta_0| > \varepsilon\} \leq c' e^{-c|\Lambda|}$$

for sufficiently large Λ, where $c, c' > 0$ are independent of Λ.

REMARKS.

1. Since $L_{\Lambda,z}(y(\Lambda),\theta)$ is continuous and Θ is compact, there always exists at least one minimizer $\hat{\theta}_{\Lambda,z}$. A measurable choice of $\hat{\theta}_{\Lambda,z}$ can be obtained by standard procedures.
2. Theorem 2.1 holds if $l_{\Lambda,z}(y(\Lambda),\theta)$ is replaced by the log-likelihood function (2.15) (see Section 3).
3. The limit $\Lambda \to Z^d$ in Theorem 2.1 and throughout the paper is taken in the sense of van Hove, that is, in the sense

$$|\Lambda| \to +\infty,$$

$$\frac{|(\Lambda + i)/\Lambda|}{|\Lambda|} \to 0 \quad \text{for every } i \in Z^d.$$

Roughly speaking, this means that the "boundary" of Λ divided by $|\Lambda|$ goes to 0 as $|\Lambda| \to +\infty$. For simplicity, the reader may assume that Λ is a hypercube of side N, so that $\Lambda \to Z^d$ means $N \to +\infty$.

4. As we mentioned in Section 1, a consistency result under stronger assumptions (Ω_0 finite, P_{θ_0} stationary, pointwise degradation) has been established in [39] by a different method. In particular, the method of [39] does not give the existence of an exponential rate c.

Next, we turn to the consistency for noncompact Θ. The following simple example, with i.i.d. random variables, demonstrates that the identifiability condition (2.16) does not suffice for proving consistency in the case of noncompact Θ. Let x_i, $i \in Z$, be independent random variables with common density

$$(e^{-2\theta} + e^{-\theta} + 2)^{-1} e^{\theta \min(x, 0)} \tag{2.17}$$

with respect to the counting measure on $\{-2, -1, 1, 2\}$. Suppose that the observed process is $y_i = |x_i|$ [i.e., $P(y|x) = \delta_{y, |x|}$]. This is easily seen to be a Bernoulli process on $\{1, 2\}$ with density such that

$$P_\theta(y = 1) = \frac{e^{-\theta} + 1}{e^{-2\theta} + e^{-\theta} + 2} \equiv g(\theta).$$

The function $g(\theta)$ achieves its maximum $\bar{g} = g(\bar{\theta})$ at $\bar{\theta} = \log(1 + \sqrt{2})$ and is strictly decreasing from $\bar{g}$ to $\frac{1}{2}$ as θ ranges over $[\bar{\theta}, +\infty)$. Take $\Theta = \{0\} \cup [\bar{\theta}, +\infty)$ and note that $g(0) = g(\infty) = \frac{1}{2}$. This model is identifiable in the sense of (2.16), but it is easily seen that

$$P_0\Big(\lim_{n \to \infty} \hat{\theta}_n = +\infty\Big) = \tfrac{1}{2}$$

and hence consistency fails.

The above example demonstrates that we need an appropriate identifiability at ∞. Our identifiability condition at ∞ involves a relative entropy and is defined as follows: Let $K_s(\theta)$ be the set of translation invariant elements in $K(\theta)$. In Corollary 3.1, we show that for any $P_{\theta_0} \in K_s(\theta_0)$, the entropy of P_{θ_0} relative to a $P_\theta \in K_s(\theta)$ defined by

$$h(P_{\theta_0}; P_\theta) = \lim_{\Lambda \to Z^d} \left\{ -\frac{1}{|\Lambda|} E_{P_\theta} \log \frac{dP_{\theta_0}^{(\Lambda)}}{dP_\theta^{(\Lambda)}} \right\} \tag{2.18}$$

exists and is independent of $P_\theta \in K_s(\theta)$ (it depends only on $\theta \in \Theta$). Furthermore, in Lemma 3.2, we show that the identifiability condition (2.16) implies that $\sup_{P_{\theta_0} \in K_s(\theta_0)} h(P_{\theta_0}; P_\theta) < 0$ for $\theta \neq \theta_0$. We will say that $\theta_0 \in \Theta$ is *identifiable at* ∞ if

$$\lim_{A \to \infty} \sup_{\theta \in \Theta:\, |\theta| \geq A} \sup_{P_{\theta_0} \in K_s(\theta_0)} h(P_{\theta_0}; P_\theta) < 0. \tag{2.19}$$

We note that this condition fails for the above example of (2.17). Condition (2.19) is natural: Intuitively, it says that the limit points of P_θ for large θ are separated from P_{θ_0} in the sense of relative entropy.

As we mentioned in Section 1, consistency for noncompact Θ will be proven under an assumption on the pressure $p_{\Lambda,z}(\theta)$. For a unit vector $\boldsymbol{\theta}$ in $\mathbb{R}^m$ (i.e., $\boldsymbol{\theta} \in \mathscr{S}^{m-1}$), we define

$$m_{\Lambda,z}(\boldsymbol{\theta}) = \max\left\{\boldsymbol{\theta} \cdot \frac{1}{|\Lambda|} U_{\Lambda,z}(x(\Lambda)) \colon x(\Lambda) \in \Omega_{\Lambda,x}\right\}. \tag{2.20}$$

ASSUMPTION. Let $A > 0$, $\theta \in \mathbb{R}^m - \{0\}$, $\boldsymbol{\theta} = \theta/|\theta|$ and

$$u_{\Lambda,z}(\theta) = p_{\Lambda,z}(\theta) - p_{\Lambda,z}(A\boldsymbol{\theta}) - |\theta - A\boldsymbol{\theta}| m_{\Lambda,Z}(\boldsymbol{\theta}).$$

We will assume that

$$\lim_{A\to\infty} \limsup_{\Lambda\to Z^d} \sup_{\theta\in\Theta\colon |\theta|\geq A} |u_{\Lambda,z}(\theta)| = 0. \tag{2.21}$$

We emphasize that condition (2.21) refers only to the MRF parametrized by θ, and it does not involve the observed process y. We also observe that $u_{\Lambda,z}(\theta)$ is nonpositive, and hence condition (2.21) amounts to a lower bound only.

Our consistency theorem for noncompact Θ is as follows.

THEOREM 2.2. *Let θ_0 and P_{θ_0} be as in Theorem 2.1, and let $\Theta \subseteq \mathbb{R}^m$ be noncompact. Assume that θ_0 is identifiable, identifiable at ∞, and that condition (2.21) holds (for a given family of b.c. z). Let $\hat{\theta}_{\Lambda,z}$ be a measurable minimizer of $l_{\Lambda,z}(y(\Lambda),\theta)$ over Θ. Then the conclusions of Theorem 2.1 hold.*

REMARKS.

1. The proof of Theorem 2.2 shows that the theorem is also true for the log-likelihood function (2.15), provided that (2.21) holds when $|u_{\Lambda,z}(\theta)|$ is replaced by $\sup_z|u_{\Lambda,z}(\theta)|$.
2. In the i.i.d. case (i.e., when P_θ is a product measure for all θ), condition (2.19) is implied by the condition

$$\left\{\bigcap_{A>0}\left[\bigcup_{\theta\in\theta\colon|\theta|\geq A} K(\theta)\right]^{\mathrm{cl}}\right\} \cap K_s(\theta_0) = \varnothing, \tag{2.22}$$

   where $[\ ]^{\mathrm{cl}}$ denotes the closure in the set of probability measures. We do not know whether the simpler condition (2.22) implies (2.19) in general. Condition (2.22) fails (as it should!) for the example of (2.17).
3. In the Appendix, we show that condition (2.21) holds if the asymptote of the finite-volume pressure $p_{\Lambda,z}(\cdot)$ converges, as Λ increases to Z^d, to the asymptote of the infinite-volume pressure $p(\cdot)$. Furthermore, we show that this is the case when $\Theta = \mathbb{R}$, $\Omega_{0,x}$ is finite, Φ is of finite range and z is the free boundary condition.
4. In addition to Theorem 2.2; we have an alternative result (to appear elsewhere) on the consistency for noncompact Θ under the assumption that the ground states of the Hamiltonian are periodic with respect to a subgroup of Z^d of finite index. This assumption holds [26] for the ferromag-

netic (but not the antiferromagnetic) Ising model. The assumption appears to be, in general, appropriate for models that satisfy the Peierls condition [34, 26].

3. A variational principle for the conditional pressure and proof of Theorem 2.1. In this section we will use the following notation: If $\mathscr{P}$ is a set of probability measures on a measurable space $\mathscr{M}$ of the form $\mathscr{M} = \mathscr{M}_0^{Z^d}$, $\mathscr{P}_s$ will denote the translation-invariant measures in $\mathscr{P}$ and $\mathscr{P}_e$ will denote the ergodic measures in $\mathscr{P}$. The set of probability measures on $\mathscr{M}$ will be denoted by $\mathscr{P}(\mathscr{M})$. If Λ is a subset of Z^d, and $R \in \mathscr{P}(\mathscr{M})$, then $R^{(\Lambda)}$ will denote the restriction of R to $\mathscr{M}_\Lambda = \mathscr{M}_0^\Lambda$. If μ_0 is a probability measure on $\mathscr{M}_0$, then we define $\mu_0^{(\Lambda)} = \mu_0^{\otimes\Lambda}$ for $\Lambda \subset Z^d$. If $R \in \mathscr{P}_s(\mathscr{M})$, then the *entropy* $h(R)$ of R (relative to μ_0) is defined by

$$h(R) = \lim_{\Lambda \to Z^d} h_\Lambda(R),$$

where

$$h_\Lambda(R) = \begin{cases} -\dfrac{1}{|\Lambda|} E_R\left(\log \dfrac{dR^{(\Lambda)}}{d\mu_0^{(\Lambda)}}\right), & \text{if } R^{(\Lambda)} \ll \mu_0^{(\Lambda)}, \\ -\infty, & \text{otherwise}. \end{cases}$$

It is well known [36] that $h(\cdot)$ is affine, upper semicontinuous in the topology of weak convergence and has compact level sets $\{R \in \mathscr{P}_s(\mathscr{M}): h(R) \geq -a\}$ for all $a \geq 0$. If $R \in \mathscr{P}(\Omega)$ (recall that $\Omega = \Omega_x \times \Omega_y = \Omega_0^{Z^d} = \Omega_{0,x}^{Z^d} \times \Omega_{0,y}^{Z^d}$), then the marginal of y will be denoted by R^y and the marginal of x will be denoted by R^x.

The proof of Theorem 2.1 will be based on Theorems 3.1 and 3.2 below. The first theorem gives a variational principle for the conditional pressure—a result of independent interest.

THEOREM 3.1. *Assume model* (2.11) *for* $P^{y|x}$. *Then*:

(i) *For any* $Q \in \mathscr{P}_s(\Omega_y)$,

$$p_{\Lambda,z}(y;\theta) \to p(y;\theta), \quad Q\text{-}a.s., \text{ as } \Lambda \to Z^d. \tag{3.1}$$

The limit $p(\cdot;\theta)$ *is independent of the b.c.* z.

(ii) *If* $Q \in \mathscr{P}_e(\Omega_y)$, *then* $p(\cdot,\theta)$ *is* Q-*a.s. a constant*, $p(Q,\theta)$, *which satisfies the following variational principle provided that* $h(Q) > -\infty$.

$$p(Q,\theta) = -h(Q) + \sup_{\substack{R\in\mathscr{P}_s(\Omega)\\ R^y = Q}} \{\theta \cdot E_R(A_U) + E_R(A_\Psi) + h(R)\}, \tag{3.2}$$

where A_U *is defined in* (2.5) *and* A_Ψ *is defined below* (2.13).

REMARKS.

1. In addition to (3.2), we can show that $p(\cdot;\theta)$ satisfies an a.s. variational principle under $Q \in \mathscr{P}_s(\Omega_y)$; this result is not needed in this paper, and we do not provide its proof here.

2. A particular case of (3.2) has been proven in [29], and for i.i.d. fields Q in [7].
3. If Q is not in $\mathscr{P}_s(\Omega_y)$, then $p_{\Lambda,z}(y,\theta)$ does in general have a limit as $\Lambda \to Z^d$. The following theorem controls the limit points of $p_{\Lambda,z}(y,\theta)$ under $P_{\theta_0} \in K(\theta_0)$.

THEOREM 3.2. *Assume model* (2.1) *for* $P^{y|x}$. *Let* $\theta_0 \in \Theta$ *and* $P_{\theta_0} \in K(\theta_0)$. *Then for any* $\varepsilon > 0$ *we have*

$$\limsup_{\Lambda \to Z^d} \frac{1}{|\Lambda|} \log P_{\theta_0}\Big\{ p_{\Lambda,z}(y,\theta) - p_{\Lambda,z}(y,\theta_0) \geq \sup_{Q \in K_e(\theta_0)} [P(Q,\theta) - p(Q,\theta_0)] + \varepsilon \Big\} < 0, \tag{3.3}$$

where $K_e(\theta_0)$ *denotes the ergodic measures in* $K(\theta_0)$.

To control the behavior of $l_{\Lambda,z}(y,\theta)$ as $\Lambda \to Z^d$, we need some properties of $p_{\Lambda,z}(\theta)$. They are given by the following well-known [36] proposition.

PROPOSITION 3.1.

(i) $p_{\Lambda,z}(\theta)$ *is convex in* θ.
(ii) $|p_{\Lambda,z}(\theta) - p_{\Lambda,z}(\theta')| \leq |\theta - \theta'|\,\|U\|$.
(iii) $|p_{\Lambda,z}(\theta)| \leq |\theta|\,\|U\|$.
(iv) *The following limit exists and is independent of* z:

$$\lim_{A \to Z^d} p_{\Lambda,z}(\theta) = p(\theta).$$

(v) *The limit* $p(\theta)$ *satisfies the variational principle*

$$p(\theta) = \sup_{R \in \mathscr{P}_s(\Omega_x)} \{\theta \cdot E_R(A_U) + h(R)\}. \tag{3.4}$$

Parts (iv) and (v) of the above proposition should be compared with parts (i) and (ii) of Theorem 3.1. The following lemma gives some basic properties for the conditional pressure, similar to properties (i)–(iii) of Proposition 3.1.

LEMMA 3.1.

(i) $P_{\Lambda,z}(y;\theta)$ *is convex in* θ (*for every* $y \in \Omega_y$).
(ii) $|p_{\Lambda,z}(y;\theta)| \leq |\theta|\,\|U\| + \|\Psi\|$, *where* $\|\Psi\|$ *is defined in* (2.16).
(iii) $|P_{\Lambda,z}(y;\theta) - p_{\Lambda,z}(y;\theta')| \leq |\theta - \theta'|\,\|U\|$.
(iv) *Let* $p_\Lambda(y;\theta)$ *be the conditional pressure with free b.c. Then for every* $\theta \in \Theta$,

$$|p_{\Lambda,z}(y;\theta) - p_\Lambda(y,\theta)| \to 0 \quad \text{as } \Lambda \to Z^d,$$

uniformly in y *and* z.

PROOF. The proofs of parts (i) and (ii) are straightforward. To prove parts (iii) and (iv), we use the following inequality: For any probability measure ν

and real functions $f, g \in L_\infty(d\nu)$, we have

$$\left|\log \int e^f \, d\nu - \log \int e^g \, d\nu\right| \le \|f - g\|_\infty. \tag{3.5}$$

Part (iii) is obtained by using (3.5) and

$$|\theta \cdot U_{\Lambda, z}(x(\Lambda)) - \theta' \cdot U_{\Lambda, z}(x(\Lambda))| = \theta - \theta'| \, |U_{\Lambda, z}(\Lambda)| \le |\theta - \theta'| \, |\Lambda| \, \|U\|.$$

To prove part (iv), we will show that

$$\begin{aligned} |U_{\Lambda, z}(x(\Lambda)) - U_\Lambda(x(\Lambda))| &\le C(\Lambda), \\ |\Psi_{\Lambda, z}(x(\Lambda), y(\Lambda)) - \Psi_\Lambda(x(\Lambda), y(\Lambda))| &\le C(\Lambda) \end{aligned} \tag{3.6a}$$

with a constant $C(\Lambda)$ satisfying

$$\frac{C(\Lambda)}{|\Lambda|} \to 0 \quad \text{as } \Lambda \to Z^d. \tag{3.6b}$$

This together with (3.5) easily yields part (iv). By (2.1),

$$\begin{aligned} U^\alpha_{\Lambda, z}(x(\Lambda)) - U^\alpha_\Lambda(x(\Lambda)) &= \sum_{V \subset Z^d} \Phi(x(V) \vee z(V)) \\ &= \sum_{i \in \Lambda} \sum_{i \in V \subset Z^d} \frac{\Phi(x(V) \vee z(V))}{|V \cap \Lambda|}, \end{aligned}$$

where Σ' denotes summation as in (2.1). Since the set $\mathscr{B}_0$ of finite-range interactions is dense in $\mathscr{B}$, we can approximate $\Phi^{(\alpha)}$, $\alpha = 1, \dots, m$, by finite-range interactions $\tilde{\Phi}^{(\alpha)}$ of interaction radius R_0. Given $\varepsilon > 0$, we can choose R_0 so that

$$\sum_{i \in V \subset Z^d \text{ finite}} \sup_{x(V)} |\Phi^{(\alpha)}(x(V)) - \tilde{\Phi}^\alpha(x(V))| < \varepsilon$$

for all $i \in Z^d$ (by translation invariance). Hence

$$|U_{\Lambda, z}(x(\Lambda)) - U_\Lambda(x(\Lambda))| \le \varepsilon |\Lambda| + 2|\partial\Lambda| \, \|U\|,$$

where $|\partial\Lambda|$ is the number of pixels which have distance from the boundary of Λ no greater than R_0. Since $|\partial\Lambda|/|\Lambda| \to 0$ as $\Lambda \to Z^d$ and ε is arbitrary, we deduce (3.6). The proof for Ψ is simpler, since Ψ has finite range. □

REMARK. Part (iv) of Lemma 3.1 and part (iv) of Proposition 3.1 show that it suffices to study the conditional pressure and the log-likelihood function with free boundary conditions only. In the rest of the paper, we consider only free boundary conditions.

PROOF OF THEOREM 3.1(i). Approximating the interactions $\Phi^{(\alpha)}$ by finite-range interactions as in the proof of Lemma 3.1, it suffices to prove the theorem for finite-range interactions only. Thus we assume that $\Phi^{(\alpha)}$ and Ψ have interaction radius r. Let $n < N$, $\Lambda_N = [-N, N]^d \subset Z^d$ and $\Lambda_n = [-n, n]^d \subset Z^d$. For each $i_0 \in [-n, n + r]^d \subset Z^d$, we consider the collection

$j(2n+r)+i_0+\Lambda_n$, $j \in Z^d$, of disjoint windows separated by corridors of width r. Each one of these disjoint windows has volume $(2n)^d$. Let $i = j(2n+r)+i_0$ and let I_{N,n,i_0} be the collection of such i's so that $i+\Lambda_n \subset \Lambda_N$. Then

$$e^{-D_{N,n}} \prod_{i \in I_{N,n,i_0}} Z_{i+\Lambda_n}(y,\theta) \le Z_{\Lambda_N}(y,\theta) \le e^{D_{N,n}} \prod_{i \in I_{N,n,i_0}} Z_{i+\Lambda_n}(y,\theta),$$

where

$$D_{N,n} = (|\theta|\,\|U\| + \|\Psi\|)\left[\left(\frac{N}{n+r}\right)^d + 2^d(n+r)N^{d-1}\right].$$

Hence

$$p_{\Lambda_N}(y,\theta) = \frac{|\Lambda_n|}{|\Lambda_N|} \sum_{i \in I_{N,n,i_0}} p_{i+\Lambda_n}(y,\theta) + O\left(\frac{1}{n^d} + \frac{n}{N}\right).$$

Now averaging over $i_0 \in [-n, n+r]^d \subset Z^d$, we obtain

$$p_{\Lambda_N}(y,\theta) = \frac{1}{(2n+1+r)^d} \frac{|\Lambda_n|}{|\Lambda_N|} \sum_{i_0 \in [-n,n+r]^d} \sum_{i \in I_{N,n,i_0}} p_{i+\Lambda_n}(y,\theta) + O\left(\frac{1}{n^d} + \frac{n}{N}\right).$$

Now, the double sum contains $(2N+1-r)^d$ terms uniformly bounded in Λ_n and y. Hence

$$p_{\Lambda_N}(y,\theta) = \frac{1}{(2N+1-r)^d} \sum_{i \in \Lambda_N:\, i+\Lambda_n \subset \Lambda_N} p_{i+\Lambda_n}(y,\theta) + O\left(\frac{1}{n^d} + \frac{n}{N} + \frac{1}{nN}\right). \tag{3.7}$$

If Q is ergodic, then by the ergodic theorem,

$$p_{\Lambda_N}(y,\theta) = E_Q\big(p_{\Lambda_n}(y,\theta)\big) + O\left(\frac{1}{n^d}\right) + R_{N,n}(y;\theta;\theta_0) \tag{3.8}$$

with

$$R_{N,n}(y;\theta;\theta_0) \to 0, \quad Q\text{-a.s., as } N \to +\infty.$$

From (3.8) we obtain for each n,

$$|P_{\Lambda_N}(y;\theta) - p_{\Lambda_{N'}}(y,\theta)| \le |R_{N,n}| + |R_{N',n}| + O\left(\frac{1}{n^d}\right),$$

which implies that $p_{\Lambda_N}(y,\theta)$ is, Q-a.s., a Cauchy sequence with some limit $p(y;\theta)$. Taking the limit $N \to +\infty$ and then $n \to +\infty$ in (3.8), we obtain

$$\lim_{N\to+\infty} p_{\Lambda_N}(y,\theta) = \lim_{n\to+\infty} E_Q\big(p_{\Lambda_n}(y,\theta)\big). \tag{3.9}$$

Since p_{Λ_n} is uniformly bounded in Λ_n and y, the limit on the right-hand side of (3.9) can be taken inside the expectation, yielding that $p(\cdot\,;\theta)$ is Q-a.s., a constant $p(Q;\theta)$.

If Q is only translation invariant (but not ergodic), then the ergodic theorem yields (1.8) with the expectation on the right-hand side of (3.8) replaced by a conditional expectation over the σ-field formed by the translation-invariant (measurable) subsets of Ω_y. The remaining steps in the proof go through and yield a random limit $p(\cdot\,;\theta)$. This completes the proof of part (i) of the theorem. □

Proof of Theorem 3.1(ii). Let $R \in \mathscr{P}_s(\Omega)$ with $h(R) > -\infty$. Then $R^{(\Lambda)}$ is absolutely continuous with respect to $\prod_{i\in\Lambda}\mu_0(dx_i, dy_i)$, and we have R-a.s.,

$$R^{(\Lambda)}(dx(\Lambda)|y(\Lambda)) = R^{(\Lambda)}(x(\Lambda)|y(\Lambda)) \prod_{i\in\Lambda}\mu_0(dx_i|y_i).$$

Therefore

$$\begin{aligned}
\theta\cdot\frac{1}{|\Lambda|}&E_R(U_\Lambda|y(\Lambda)) + \frac{1}{|\Lambda|}E_R(\Psi_\Lambda|y(\Lambda)) + \frac{1}{|\Lambda|}E_R[\log R^{(\Lambda)}(x(\Lambda)|y(\Lambda))|y(\Lambda)]\\
&= \frac{1}{|\Lambda|}E_R\left\{\log\frac{e^{\theta\cdot U_\Lambda(x(\Lambda))+\Psi_\Lambda(x(\Lambda),y(\Lambda))}}{R^{(\Lambda)}(x(\Lambda)|y(\Lambda))}\middle|y(\Lambda)\right\}\\
&\le \frac{1}{|\Lambda|}\log\int e^{\theta\cdot U_\Lambda(x)+\Psi_\Lambda(x,y)}\prod_{i\in\Lambda}\mu_0(dx_i|y_i)\\
&= p_\Lambda(y,\theta).
\end{aligned}$$

Hence

$$\begin{aligned}
p_\Lambda(y,\theta) &- \frac{1}{|\Lambda|}\log Q^{(\Lambda)}(y)\\
&\ge \frac{1}{|\Lambda|}E_R(\theta\cdot U_\Lambda + \Psi_\Lambda|y(\Lambda)) - \frac{1}{|\Lambda|}E_R\{\log R^{(\Lambda)}(x|y)Q^{(\Lambda)}(y)|y(\Lambda)\}.
\end{aligned}$$

Here and below, we write $Q^{(\Lambda)}(y)$ for the derivative of $Q^{(\Lambda)}$ with respect to $\prod_{i\in\Lambda}\mu_{0,y}(dy_i)$.

Assuming $R^y = Q$ and integrating, we obtain

$$\begin{aligned}
&E_Q(p_\Lambda(y,\theta)) - \frac{1}{|\Lambda|}E_Q(\log Q^{(\Lambda)})\\
&\qquad\ge \frac{1}{|\Lambda|}E_R(\theta\cdot U_\Lambda + \Psi_\Lambda) - \frac{1}{|\Lambda|}E_R(\log R^{(\Lambda)}).
\end{aligned}\tag{3.10}$$

Since R is translation invariant, the right-hand side of (3.10) converges to

$$E_R(\theta\cdot A_U + A_\Psi) + h(R).$$

By part (i) of the theorem, $E_Q(p_\Lambda(y,\theta))$ converges to $p(Q,\theta)$. Hence we obtain

$$p(Q,\theta) + h(Q) \ge \theta\cdot E_R(A_U) + E_R(A_\Psi) + h(R).\tag{3.11}$$

In particular,

$$(3.12)\quad p(Q,\theta) + h(Q) \geq \sup_{\substack{R \in \mathscr{P}(\Omega) \\ R^y = Q}} \{Q \cdot E_R(A_U) + E_R(A_\Psi) + h(R)\}.$$

To prove that $p(Q,\theta) + h(Q)$ is actually equal to the supremum in (3.12), we will use an explicit construction. Let

$$d\rho_\Lambda(x,y) = \frac{e^{\theta \cdot U_\Lambda(x) + \Psi_\Lambda(x,y)}}{Z_\Lambda(y,\theta)} \prod_{i \in \Lambda} \mu_0(dx_i|y_i) Q^{(\Lambda)}(dy(\Lambda)).$$

Note that $\rho_\Lambda \in \mathscr{P}(\Omega_0^\Lambda)$ and

$$(3.13)\quad \begin{aligned} &E_Q(p_\Lambda(y,\theta)) - \frac{1}{|\Lambda|} E_Q(\log Q^{(\Lambda)}) \\ &= \frac{1}{|\Lambda|} E_{\rho_\Lambda}(\theta \cdot U_\Lambda) + \frac{1}{|\Lambda|} E_{\rho_\Lambda}(\Psi_\Lambda) - \frac{1}{|\Lambda|} E_{\rho_\Lambda}(\log \rho_\Lambda). \end{aligned}$$

Assuming that Λ is a hypercube of side N, we construct a $\bar{\rho}_\Lambda \in \mathscr{P}(\Omega)$ by taking translates of ρ_Λ in each hypercube $\Lambda + jN$, $j \in Z^d$, and then defining $\bar{\rho}_\Lambda$ as the product of these translates. The probability measure $\bar{\rho}_\Lambda$ is periodic but not translation invariant. To obtain a translation-invariant distribution, we average over Λ, that is, we define

$$\hat{\rho}_\Lambda = \frac{1}{|\Lambda|} \sum_{i \in \Lambda} \bar{\rho}_\Lambda \circ \tau^i.$$

Since $\bar{\rho}_\Lambda$ is an i.i.d. field on $(\Omega_0^\Lambda)^{Z^d}$, its entropy relative to μ_0,

$$h(\bar{\rho}_\Lambda) = - \lim_{\Lambda' \to Z^d} \frac{1}{|\Lambda'|} E_{\bar{\rho}_\Lambda}(\log \bar{\rho}_\Lambda^{(\Lambda')}),$$

is well defined and equal to $(1/|\Lambda|)E_{\rho_\Lambda}(\log \rho_\Lambda)$. Furthermore,

$$h(\bar{\rho}_\Lambda \circ \tau^i) = - \frac{1}{|\Lambda|} E_{\rho_\Lambda}(\log \rho_\Lambda)$$

for all $i \in Z^d$. By the linearity of the entropy, we have

$$h(\hat{\rho}_\Lambda) = h(\bar{\rho}_\Lambda) = - \frac{1}{|\Lambda|} E_{\rho_\Lambda}(\log \rho_\Lambda).$$

Using the same procedure as in the proof of (3.6), one can easily show that

$$E_{\hat{\rho}_\Lambda}(\theta \cdot A_U + A_\Psi) = \frac{1}{|\Lambda|} \sum_{i \in \Lambda} \int (\theta \cdot A_U + A_\Psi) \circ \tau^i \, d\bar{\rho}_\Lambda$$

and

$$\frac{1}{|\Lambda|} E_{\rho_\Lambda}(\theta \cdot U_\Lambda + \Psi_\Lambda) = \frac{1}{|\Lambda|} \sum_{i \in \Lambda} \sum_{i \in V \subset \Lambda} \int \left[\frac{\theta \cdot \Phi(x(V)) + \Psi(x,y)}{|V|} \right] d\rho_\Lambda$$

differ by $\varepsilon(|\Lambda|)$. Hence

$$\lim_{\Lambda \to Z^d} \left\{ E_Q(p_\Lambda(y,\theta)) - \frac{1}{|\Lambda|} E_Q(\log Q^{(\Lambda)}) \right\} = \lim_{\Lambda \to Z^d} \left\{ E_{\hat{\rho}_\Lambda}(\theta \cdot A_U + A_\Psi) + h(\hat{\rho}_\Lambda) \right\},$$

that is,

$$p(Q,\theta) + h(Q) = \lim_{A \to Z^d} \left\{ E_{\hat{\rho}_\Lambda}(\theta \cdot A_U + A_\Psi) + h(\hat{\rho}_\Lambda) \right\}. \tag{3.14}$$

This, together with (3.12), implies that

$$\liminf_{\Lambda \to Z^d} h(\hat{\rho}_\Lambda) > -\infty.$$

Thus $\{\hat{\rho}_\Lambda\}$ is a tight sequence in $\mathscr{P}_s(\Omega)$. Also note that the marginal of y, $(\hat{\rho}_\Lambda)^y$, converges to Q as $\Lambda \to Z^d$. By (3.14) any limit point ρ of $\hat{\rho}_\Lambda$ achieves equality in (3.12). This completes the proof of the theorem. □

The following proposition will be used in the proof of Theorem 3.2. It is based on a large deviation result for MRFs [6, 13, 32].

PROPOSITION 3.2. *Assume model* (2.11) *for* $P^{y|x}$. *Let*

$$R_{\Lambda,y} = \frac{1}{|\Lambda|} \sum_{i \in \Lambda} \delta_{\tau^i y}$$

be the empirical field of the observed process in a finite window $\Lambda \subset Z^d$. *Let*

$$F \colon \mathscr{P}(\Omega_y) \to \mathbb{R}$$

be a lower semicontinuous function on $\mathscr{P}(\Omega_y)$. *Then for any* $P_{\theta_0} \in K(\theta_0)$ *and all* $\varepsilon > 0$, *we have*

$$\limsup_{\Lambda \to Z^d} \frac{1}{|\Lambda|} \log P_{\theta_0} \left\{ F(R_{\Lambda,y}) \le \min_{Q \in K_s(\theta_0)} F(Q) - \varepsilon \right\} < 0. \tag{3.15}$$

PROOF. Our assumptions on $P^{y|x}$ imply that the pair (x,y) is an MRF. Let

$$R_{\Lambda,(x,y)} = \frac{1}{|\Lambda|} \sum_{i \in \Lambda} \delta_{\tau^i(x,y)}$$

be the empirical field of (x,y). Since (x,y) is an MRF, we have [6, 13, 32]

$$\limsup_{\Lambda \to Z^d} \frac{1}{|\Lambda|} \log \tau_{\theta_0} \otimes P^{y|x} \left\{ \tilde{F}(R_{\Lambda,(x,y)}) \le \min_{R \in \tilde{G}_s(\theta_0)} \tilde{F}(R) - \varepsilon \right\} < 0 \tag{3.16}$$

for any $\pi_{\theta_0} \in G(\theta_0)$, and all $\varepsilon > 0$. Here $\tilde{F}$ is a real-valued lower-semicontinuous function on $\mathscr{P}_s(\Omega)$, $\Omega = \Omega_x \times \Omega_y$ and $\tilde{G}_s(\theta_0)$ is the set of stationary Gibbs distributions for the interaction $\theta \cdot \Phi + \Psi$. We claim that

$$\tilde{G}_s(\theta_0) = G_s(\theta_0) \otimes P^{y|x}.$$

Indeed, $G(\theta_0) \otimes P^{y|x} \subset \tilde{G}(\theta_0)$ by the first remark below (2.13). On the other hand, for any $R \in \tilde{G}(\theta_0)$, using

$$\int e^{\Psi_0(x(W), y_0)} \mu_0(dy_0|x_0) = 1$$

and (2.8), we see that $R(dx)$ itself satisfies (2.8) with interactions $\theta_0 \cdot \Phi$ and that $R(dy|x) = P^{y|x}$. Hence $R = R^x \otimes p^{y|x} \in G(\theta_0) \otimes P^{y|x}$ and therefore $\tilde{G}(\theta_0) = G(\theta_0) \otimes P^{y|x}$. This yields the claim. □

Now, for a given F as above, we define an $\tilde{F}$ on $\mathscr{P}_s(\Omega)$ so that $\tilde{F}(R) = F(R^y)$. Then $\tilde{F}(R_{\Lambda,(x,y)}) = F(R_{\Lambda,y})$. Hence (3.16) becomes

$$\text{(3.17)} \qquad \limsup_{R \to Z^d} \frac{1}{|\Lambda|} \log P_{\theta_0}\left\{ F(R_{\Lambda,y}) \le \min_{R \in \tilde{G}_s(\theta_0)} F(R^y) - \varepsilon \right\} < \varepsilon.$$

Now, it is easily seen that

$$\min_{R \in \tilde{G}_s(\theta_0)} F(R^y) = \min_{Q \in K_s(\theta_0)} F(Q).$$

This together with (3.17) yields (3.15). □

PROOF OF THEOREM 3.2. Let Λ_N, Λ_n be as in the proof of Theorem 3.1(i). Let $\tilde{\Lambda}_N = [-N, N-r]^d \subset Z^d$ and consider the empirical field

$$R_{\tilde{\Lambda}_{N,y}} = \frac{1}{|\tilde{\Lambda}_N|} \sum_{i \in \tilde{\Lambda}_N} \delta_{\tau^i y}.$$

Then, using (3.7), we obtain

$$\text{(3.18)} \qquad p_{\Lambda_N}(y;\theta) = \int p_{\Lambda_n}(y',\theta) R_{\tilde{\Lambda}_{N,y}}(dy') + O\left(\frac{1}{n^d} + \frac{1}{N} + \frac{1}{nN}\right)$$

and a similar expression for $p_{\Lambda_N}(y, \theta_0)$. For a fixed n we define

$$f(y) = p_{\Lambda_n}(y;\theta) - p_{\Lambda_n}(y, \theta_0)$$

and

$$F(R) = -\int f\,dR$$

for any $R \in \mathscr{P}_s(\Omega_y)$. The function f is a continuous function on Ω_y, and by part (ii) of Lemma 3.1 it is bounded. Hence F is a bounded continuous function on $\mathscr{P}(\Omega_y)$. By (3.15), we have (since $|\tilde{\Lambda}_N|/|\Lambda_N| \to 1$ as $N \to +\infty$):

$$\text{(3.19)} \qquad \begin{aligned} &\limsup_{N \to +\infty} \frac{1}{|\Lambda_N|} \log P_{\theta_0}\left\{ \int \left[p_{\Lambda_n}(y',\theta) - p_{\Lambda_n}(y',\theta_0) \right] R_{\tilde{\Lambda}_{n,y}}(dy') \right. \\ &\qquad \left. \ge \max_{Q \in K_s(\theta_0)} \int \left[p_{\Lambda_n}(y',\theta) - p_{\Lambda_n}(y',\theta_0) \right] dQ(y') + \frac{\varepsilon}{4} \right\} < 0. \end{aligned}$$

Now from Theorem 3.1, we have for Q ergodic:

$$\lim_{n\to+\infty} E_Q\big[p_{\Lambda_n}(y,\theta) - p_{\Lambda_n}(y,\theta_0)\big] = p(Q,\theta) - p(Q,\theta_0),$$

and hence

$$\liminf_{n\to+\infty} \max_{Q\in K_s(\theta_0)} E_Q\big[p_{\Lambda_n}(y,\theta) - p_{\Lambda_n}(y,\theta_0)\big] \ge \sup_{Q\in K_e(\theta_0)} \big[p(Q,\theta) - (Q,\theta_0)\big].$$

This, together with (3.19) and (3.18), yields (3.3). □

Theorem 3.1 has the following corollary.

COROLLARY 3.1.

(i) *If* $P_{\theta_0} \in K_e(\theta_0)$, *then* P_{θ_0}-*a.s.*,

$$l_\Lambda(y,\theta) - l_\Lambda(y,\theta_0)$$

$$\to p(\theta) - \sup_{\substack{R\in\mathscr{P}_s(\Omega)\\ R^y = P_{\theta_0}}} \big[\theta\cdot E_R(A_U) + E_R(A_\Psi) + h(R)\big] \quad \text{as } \Lambda\to Z^d. \tag{3.20}$$

(ii) *If* $P_{\theta_0} \in K_s(\theta_0)$, *then the relative entropy* $h(P_{\theta_0}; P_\theta)$, *given by* (2.18), *exists and is the negative of the limit in* (3.20) *when* $P_{\theta_0} \in K_e(\theta_0)$.

PROOF. We have P_{θ_0}-a.s.,

$$\begin{aligned}
\lim_{\Lambda\to Z^d} &(l_\Lambda(y,\theta) - l_\Lambda(y,\theta_0))\\
&= p(\theta) - p(P_{\theta_0},\theta) - p\big[(\theta_0) - p(P_{\theta_0},\theta_0)\big]\\
&= p(\theta) - \sup_{\substack{R\in\mathscr{P}_s(\Omega)\\ R^y = P_{\theta_0}}} \big[\theta\cdot E_R(A_U) + E_R(A_\Psi) + h(R)\big]\\
&\qquad -\big[p(\theta_0) - p(P_{\theta_0},\theta_0) - h(P_{\theta_0})\big].
\end{aligned}$$

Now

$$\begin{aligned}
p(\theta_0) &- p(P_{\theta_0},\theta_0) - h(P_{\theta_0})\\
&= p(\theta_0) - \sup_{\substack{R\in\mathscr{P}s(\Omega)\\ R^y = P_{\theta_0}}} \big[\theta_0\cdot E_R(A_U) + E_R(A_\Psi) + h(R)\big].
\end{aligned} \tag{3.21}$$

Next note that the pressure $p(\theta_0)$ is also the pressure of the MRF (x,y) whose energy function is $\theta_0\cdot U_\Lambda + \Psi_\Lambda$. Indeed, the pressure of (x,y) is $\tilde{p}(\theta_0) = \lim \tilde{p}_\Lambda(\theta_0)$ with

$$\begin{aligned}
\tilde{p}_\Lambda(\theta_0) &= \frac{1}{|\Lambda|}\log\int \exp\{\theta_0\cdot U_\Lambda(x(\Lambda)) + \Psi_\Lambda(x(\Lambda),y(\Lambda))\}\mu_0^{(\Lambda)}(dx(\Lambda),dy(\Lambda))\\
&= \frac{1}{|\Lambda|}\log\int \exp\{\theta_0\cdot U_\Lambda(x(\Lambda))\}\mu_{0,x}^{(\Lambda)}(dx(\Lambda)) = p_\Lambda(\theta_0).
\end{aligned}$$

Thus by the variational formula (3.4) we have

$$p(\theta_0) = \sup_{R \in \mathscr{P}_s(\Omega)} \left[\theta_0 \cdot E_R(A_U) + E_R(A_\Psi) + h(R)\right]. \tag{3.22}$$

The supremum in (3.22) is achieved by the stationary MRF for (x, y), that is, by $\pi_{\theta_0} \otimes P^{y|x}$ with $\pi_{\theta_0} \in G_s(\theta_0)$. Hence the supremum in (3.21) is also achieved by the stationary MRF for (x, y), and therefore the right-hand side of (3.21) is 0, that is,

$$p(P_{\theta_0}, \theta_0) = p(\theta_0) + h(P_{\theta_0}). \tag{3.23}$$

This yields part (i) of the corollary. Next we prove part (ii). Let $P_{\theta_0} \in K_e(\theta_0)$. By part (i) and Lebesgue's theorem, we have

$$\begin{aligned} h(P_{\theta_0}; P_\theta) &= -\lim_{\Lambda \to Z^d} \frac{1}{|\Lambda|} E_{P_{\theta_0}}\left\{\log \frac{dP_{\theta_0}^{(\Lambda)}}{dP_\theta^{(\Lambda)}}\right\} \\ &= -\lim_{\Lambda \to Z^d} E_{P_{\theta_0}}\{l_\Lambda(y, \theta) - l_\Lambda(y, \theta_0)\} \\ &= E_{P_{\theta_0}}\left\{\lim_{\Lambda \to Z^d} \left[l_\Lambda(y, \theta) - l_\Lambda(y, \theta_0)\right]\right\}. \end{aligned}$$

Since $P_{\theta_0} \in K_s(\theta_0)$, we can find a stationary $\pi_{\theta_0} \in G_s(\theta_0)$ such that $(\pi_{\theta_0} \otimes P^{y|x})^y = P_{\theta_0}$. Since $\pi_{\theta_0} \otimes P^{y|x}$ is itself a stationary Gibbs measure, we can decompose it into ergodic Gibbs measures. Taking the marginal on Ω_y, we obtain

$$P_{\theta_0} = \int_{Q \in K_e(\theta_0)} Q\alpha(dQ)$$

with some probability measure α on $K_e(\theta_0)$. Proceeding as above, we obtain the existence of

$$h(P_{\theta_0}; P_\theta) = \int_{K_e(\theta_0)} h(Q; P_\theta)\alpha(dQ). \tag{3.24}$$

This completes the proof of the corollary. □

The following lemma will be combined with Theorem 3.2 to prove Theorem 2.1.

LEMMA 3.2. *Let*

$$\Delta(\theta_0, \theta) = p(\theta) - \sup_{\substack{R \in \mathscr{P}_s(\Omega) \\ R^y \in K_e(\theta_0)}} \left[\theta \cdot E_R(A_U) + E_R(A_\Psi) + h(R)\right]. \tag{3.25}$$

Then:

(i) *For all* $Q \in K_e(\theta_0)$, *we have*

$$p(\theta) - p(Q, \theta) - \left[p(\theta_0) - p(Q, \theta_0)\right] \geq \Delta(\theta_0, \theta).$$

(ii) $\Delta(\theta_0, \theta) \geq 0$ *with equality iff* $\theta = \theta_0$.

(iii) $\Delta(\theta_0, \theta)$ *is continuous in* θ.

PROOF. (i) By Corollary 3.1,

$$\begin{aligned} p(\theta) - p(Q,\theta) - p(\theta_0) + p(Q,\theta_0) &= p(\theta) - \sup_{\substack{R\in\mathscr{P}_s(\Omega)\\ R^y=Q}} [\theta\cdot E_R(A_U) + E_R(A_\Psi) + h(R)] \\ &\ge \Delta(\theta_0,\theta). \end{aligned}$$

(ii) Using the variational formula (3.22) for $p(\theta)$, we obtain that $\Delta(\theta_0,\theta) \ge 0$. Now suppose that $\Delta(\theta_0,\theta) = 0$. First note that

$$F(R) = \theta\cdot E_R(A_U) + E_R(A_\Psi) + h(R)$$

is upper semicontinuous, bounded from above, and the level sets $\{R;\, F(R)\ge a\}$ are compact and nonempty for small enough a. Hence the supremum

$$\sup_{\substack{R\in\mathscr{P}(\Omega)\\ R^y\in K_e(\theta_0)}} F(R)$$

is achieved. This together with the remarks below (3.22) [applied to $p(\theta)$] imply that $\Delta(\theta_0,\theta) = 0$ iff there exists $R^x \in G_s(\theta)$ such that $(R^x \otimes p^{y|x})^y \in K_e(\theta_0)$. Therefore, $K(\theta)\cap K(\theta_0) \ne 0$. By our identifiability condition, this happens only if $\theta = \theta_0$.

(iii) $p(\theta)$ is continuous by Proposition 3.1. Also, from the definition of $\Delta(\theta_0,\theta)$, we have that $\Delta(\theta_0,\theta)$ is upper semicontinuous. It remains to prove that it is lower semicontinuous. Now, for some $R\in\mathscr{P}_s(\Omega)$ with $R^y\in K_e(\theta_0)$, we have

$$\Delta(\theta_0,\theta) - p(\theta) = -\theta\cdot E_R(A_U) - E_R(A_\Psi) - h(R).$$

Let $\theta_n \to \theta$. For some sequence $R_n \in \mathscr{P}_s(\Omega)$ with $R_n^y \in K_e(\theta_0)$, we have

$$\Delta(\theta_0,\theta_n) - p(\theta_n) = -\theta_n\cdot E_{R_n}(A_U) - E_{R_n}(A_\Psi) - h(R_n).$$

Since $h(R_n)$ is bounded, the sequence R_n is relatively compact and

$$\begin{aligned} \liminf_{n\to+\infty}[\Delta(\theta_0,\theta_n) - p(\theta_n)] &\ge -E_R(\theta\cdot A_U) - E_R(A_\Psi) - h(R) \\ &\ge \Delta(\theta_0,\theta) - p(\theta) \end{aligned}$$

for some limit point R of R_n, since $h(R)$ is lower semicontinuous and $R^y \in K_e(\theta_0)$. This proves the lower semicontinuity of $\Delta(\theta_0;\cdot)$. □

PROOF OF THEOREM 2.1. We write

$$\begin{aligned} &l_\Lambda(y,\theta) - l_\Lambda(y,\theta_0) - p_\Lambda(\theta) - p_\Lambda(y,\theta) - p_\Lambda(\theta_0) + p_\Lambda(y,\theta_0) \\ &\quad = p_\Lambda(\theta) - p(\theta) - [p_\Lambda(\theta_0) - p(\theta_0)] \\ &\qquad - [p_\Lambda(y,\theta) - p_\Lambda(y,\theta_0)] + \sup_{Q\in K_e(\theta_0)} [p(Q,\theta) - p(Q,\theta_0)] \qquad (3.26a) \\ &\qquad + p(\theta) - p(\theta_0) - \sup_{Q\in K_e(\theta_0)} [p(Q,\theta) - p(Q,\theta_0)]. \end{aligned}$$

Let D be an open neighborhood of θ_0. Since $\Delta(\theta_0, \theta)$ is continuous in θ (by Lemma 3.2), its minimum on the compact set Θ/D is achieved, and by part (ii) of Lemma 3.2, this minimum is strictly greater than 4ε for some $\varepsilon > 0$. Proposition 3.1 implies, by standard finite covering arguments, that $p_\Lambda(\theta) \to p(\theta)$ uniformly in θ, and hence

$$p_\Lambda(\theta) - p(\theta) \geq -\varepsilon$$

for sufficiently large Λ and all $\theta \in \Theta/D$. Also, for large Λ,

$$p(\theta_0) - p_\Lambda(\theta_0) \geq -\varepsilon.$$

By Lemma 3.1(iii), the family $\{p_\Lambda(y;\theta)\}$ is uniformly equicontinuous in Θ/D. This together with Theorem 3.2 imply (again by finite covering arguments)

$$(3.26\text{b}) \qquad \begin{aligned} P_{\theta_0}\Big\{ &\sup_{\theta\in\Theta/D} [p_\Lambda(y,\theta) - p_\Lambda(y,\theta_0)] \\ &- \sup_{Q\in K_e(\theta_0)} [p(Q,\theta) - p(Q,\theta_0)] \geq \varepsilon\Big\} \leq c'e^{-C|\Lambda|} \end{aligned}$$

for sufficiently large Λ and some $C, c' > 0$. The last term in (3.26a) is bounded below by $\Delta(\theta_0, \theta)$ which is larger than 4ε for $\theta \in \Theta/D$. The above lower bounds and (3.26a) yield

$$P_{\theta_0}\Big\{ \inf_{\theta\in\Theta/D} l_\Lambda(y,\theta) - l_\Lambda(y,\theta_0) \geq \varepsilon\Big\} \leq c'e^{-C|\Lambda|}$$

or equivalently

$$P_{\theta_0}\Big\{ \inf_{\theta\in\Theta/D} l_\Lambda(y,\theta) \geq \inf_{\theta'\in\Theta} l_\Lambda(y,\theta') + \varepsilon\Big\} \geq 1 - c'e^{-C|\Lambda|}.$$

But this is true for all neighborhoods D of θ_0. Taking a countable family of neighborhoods shrinking to θ_0 and applying the Borel–Cantelli lemma, we easily deduce the theorem. □

The following lemma implies that Theorem 2.1 holds also for the log-likelihood function (2.14).

LEMMA 3.3.

$$|l_\Lambda(y,\theta) - \tilde{l}_\Lambda(y,\theta)| \leq 2\frac{C(\Lambda)}{|\Lambda|}$$

with a constant $C(\Lambda)$ *satisfying*

$$(3.27) \qquad \frac{C(\Lambda)}{|\Lambda|} \to 0 \quad \text{as } \Lambda \to Z^d.$$

PROOF. Suppose that $P_\theta \in K(\theta)$ corresponds to $\pi_\theta \in G(\theta)$, that is, P_θ is the marginal of y under $\pi_\theta \otimes P^{y|x}$. Then

$$P_\theta^{(\Lambda)}(dy(\Lambda)) = \int_{\Omega_{W(\Lambda),x}} \pi_\theta^{W(\Lambda)}(dx(\Lambda))P^{y|x}(dy(\Lambda)/x(W(\Lambda))).$$

Now the fact that $\pi_\theta \in G(\theta)$ implies that for all finite $\Lambda \subset Z^d$, we have

$$\pi_\theta^{(\Lambda)}(dx(\Lambda)) = \mu_{0,x}^{(\Lambda)}(dx(\Lambda)) \int_{\Omega_{\Lambda^C,x}} \pi_{\Lambda,\theta,x(\Lambda^C)}(x(\Lambda))\, d\pi_\theta(x(\Lambda^C)).$$

Hence

$$\begin{aligned} P_\theta^{(\Lambda)}(dy(\Lambda)) &= \int_{\Omega_{W(\Lambda),x}} \mu_{0,x}^{W(\Lambda)}(dx(\Lambda)) p^{y|x}(dy(\Lambda)|x(W(\Lambda))) \\ &\quad \times \int_{\Omega_{W(\Lambda)^C,x}} \pi_{\Lambda,\theta,x(W(\Lambda^C))}(x(\Lambda))\, d\pi_\theta\big(x\big(W(\Lambda)^C\big)\big) \\ &= \mu_{0,y}^{(\Lambda)}(dy(\Lambda)) \int_{\Omega_{\Lambda,x}} \prod_{i\in\Lambda} \mu_0(dx_i|y_i) e^{\Psi_{\Lambda,x(\Lambda^C)}(x(\Lambda),y(\Lambda))} \\ &\quad \times \int_{\Omega_{W(\Lambda)C,x}} \pi_{\Lambda,\theta,x(W(\Lambda)^C)}(x(\Lambda))\, d\pi_\theta\big(x\big(W(\Lambda)^C\big)\big). \end{aligned} \tag{3.28}$$

The proof of (3.6) may be used to show that (see [23], Lemma 3.3)

$$e^{-2C(\Lambda)} \le \frac{\pi_{\Lambda,\theta,x(\Lambda^C)}(x(\Lambda))}{\pi_{\Lambda,\theta}(x(\Lambda))} \le e^{2C(\Lambda)}$$

with a constant $C(\Lambda)$ satisfying (3.27). This, together with (3.28), yields the lemma. □

4. Proof of Theorem 2.2. The proof of Theorem 2.2 uses the following lemma.

LEMMA 4.1. *If condition* (2.21) *holds, then*

$$\begin{aligned} \lim_{A\to\infty} \liminf_{\Lambda\to Z^d} \Big\{ &\inf_{|\theta|\ge A} [l_{\Lambda,z}(y,\theta) - l_{\Lambda,z}(y,\theta_0)] \\ &- \inf_{\theta\in\Theta:|\theta|=A} [l_{\Lambda,z}(y,\theta) - l_{\Lambda,z}(y,\theta_0)] \Big\} = 0. \end{aligned} \tag{4.1}$$

PROOF. By (2.14) we have

$$l_{\Lambda,z}(y,\theta) - l_\Lambda(y,A\boldsymbol{\theta}) = p_{\Lambda,z}(\theta) - p_{\Lambda,z}(A\boldsymbol{\theta}) - \big[p_{\Lambda,z}(y,\theta) - p_{\Lambda,z}(y,A\boldsymbol{\theta})\big].$$

Using the definition of the conditional pressure and of $m_{\Lambda,z}(\boldsymbol{\theta})$, one easily obtains

$$p_{\Lambda,z}(y,\theta) - p_{\Lambda,z}(y,A\boldsymbol{\theta}) \le |(\theta)| - A|m_{\Lambda,z}(\boldsymbol{\theta}).$$

Hence

$$l_{\Lambda,z}(y,\theta) \ge l_{\Lambda,z}(y,A\boldsymbol{\theta}) + u_{\Lambda,z}(\theta).$$

Therefore,

$$\inf_{|\theta|=A} l_{\Lambda,z}(y,\theta) \geq \inf_{|\theta|\geq A} l_{\Lambda,z}(y,\theta)$$
$$\geq \inf_{|\theta|=A} l_{\Lambda,z}(y,\boldsymbol{\theta}) - \sup_{|\theta|\geq A} |u_{\Lambda,z}(\theta)|,$$

which proves the lemma. □

PROOF OF THEOREM 2.2. By Lemma 3.2, the condition of identifiability at ∞ (2.19) gives

$$\liminf_{A\to\infty} \inf_{\theta\in\Theta:\,|\theta|=A} \Delta(\theta_0,\theta) \equiv \overline{\Delta}(\theta_0) > 0.$$

Let $\delta < \frac{1}{4}\overline{\Delta}(\theta_0)$ and $A = A(\delta)$ such that

$$\inf_{|\theta|=A} \Delta(\theta_0,\theta) \geq \overline{\Delta}(\theta_0) - \delta$$

and

$$\inf_{|\theta|\geq A} [l_{\Lambda,z}(y,\theta) - l_{\Lambda,z}(y,\theta_0)] \geq \inf_{|\theta|=A} [l_{\Lambda,z}(y,\theta) - l_{\Lambda,z}(y,\theta_0)] + \delta$$

for large enough Λ. These together with (3.26b) yield

$$P_{\theta_0}\Big\{ \inf_{|\theta|\geq A} [l_\Lambda(y,\theta) - l_\Lambda(y,\theta_0)] \leq \overline{\Delta}(\theta_0) - 3\delta \Big\} \leq c'e^{-c|\Lambda|}$$

for large Λ and some $c, c' > 0$. Thus

$$P_{\theta_0}\{|\hat{\theta}_{\Lambda,z}| \geq A\} \leq c'e^{-c|\Lambda|}.$$

This yields Theorem 2.2. □

APPENDIX

In this appendix, we elaborate on condition (2.21), prove it in a special case and argue that consistency in the noncompact case is related to the notion of *ground random fields* (see [20], page 454, and references cited therein). Throughout this appendix we assume that $\Omega_{0,x}$ is finite.

Let $\boldsymbol{\theta}$ be a unit vector in $\mathbb{R}^m$ (i.e., $\boldsymbol{\theta} \in s^{m-1}$). A probability measure $\pi_{\boldsymbol{\theta}}$ in $\mathscr{P}(\Omega_x)$ is said to be a ground random field (GRF) relative to the interactions $\Phi^{(\alpha)}$, $a = 1,\dots,m$ (see Section 2), and with parameter vector $\boldsymbol{\theta} \in S^{m-1}$, if for every finite $\Lambda \subset Z^d$, the density of the conditional probability distribution

$$\pi_{\boldsymbol{\theta}}(dx(\Lambda)|x(\Lambda^c))$$

is uniform on the (finite) set of configurations $x(\Lambda)$ maximizing $\boldsymbol{\theta}\cdot U_{\Lambda,x(\Lambda^c)}(x(\Lambda))$. Intuitively, this means that $\pi_{\boldsymbol{\theta}}$ satisfies (2.8) with $\theta = |\theta|\boldsymbol{\theta}$ and $|\theta| = +\infty$. An *attainable ground random field* (AGRF) is a weak limit of a sequence $\pi_{\theta_n} \in G(|\theta_n|\boldsymbol{\theta})$ as $|\theta_n| \to \infty$. For a fixed set of interactions $\Phi^{(\alpha)}$, $\alpha = 1,\dots,m$, the set of GRFs associated with a $\boldsymbol{\theta} \in S^{m-1}$ will be denoted by $G(\boldsymbol{\theta})$, and the set of AGRFs will be denoted by $G_a(\boldsymbol{\theta})$.

The set $G(\boldsymbol{\theta})$ of GRFs contains [20] the set $G_a(\boldsymbol{\theta})$ of AGRFs, and there are examples [20] for which $G_a(\boldsymbol{\theta})$ is a strict subset of $G(\boldsymbol{\theta})$. From the point of view of estimation, a condition like (2.19) [or (2.22)] controls only the set of AGRFs (and hence the corresponding distribution on Ω_y); on the other hand, the ML estimators $\hat{\theta}_{\Lambda, z}$ involve the entire set $G(\boldsymbol{\theta})$ of GRFs. This indicates that in addition to the control implied by (2.19), we need, for large Λ, an estimate which is uniform in θ for θ in the one-point compactification of $\mathbb{R}^m$. Condition (2.21) provides such an estimate.

In the rest of this appendix, we will assume that the $\Phi^{(\alpha)}$'s have finite range. Also, for simplicity we will consider free boundary conditions and we will drop the index z. For $\theta \in \mathbb{R}^m - \{0\}$, we write $\theta = |\theta|\boldsymbol{\theta}$. Let

$$g_\Lambda(\theta) = g_\Lambda(|\theta|, \boldsymbol{\theta}) = p_\Lambda(\theta) - |\theta| m_\Lambda(\boldsymbol{\theta}) \tag{A.1}$$

and

$$g(\theta) = g(|\theta|, \boldsymbol{\theta}) = p(\theta) - |\theta| m(\boldsymbol{\theta}), \tag{A.2}$$

where

$$m(\boldsymbol{\theta}) = \lim_{\Lambda \to Z^d} m_\Lambda(\boldsymbol{\theta}).$$

This limit clearly exists, since the $\Phi^{(\alpha)}$'s have finite range.

Differentiating with respect to $|\theta|$ and using the definition of $m_\Lambda(\boldsymbol{\theta})$, it is easily seen that $g_\Lambda(|\theta|, \boldsymbol{\theta})$, and hence $g(|\theta|, \boldsymbol{\theta}) = \lim_\Lambda g_\Lambda(|\theta|, \boldsymbol{\theta})$, is nonincreasing in $|\theta|$. Let

$$a = \min_{x \in \Omega_{0,x}} \mu_{0,x}\{x\}.$$

Since $g_\Lambda(|\theta|, \boldsymbol{\theta})$ and $g(|\theta|, \boldsymbol{\theta})$ are bounded below by $\log a > -\infty$, the following limits exist:

$$\lim_{|\theta| \to \infty} g_\Lambda(|\theta|, \boldsymbol{\theta}) = f_\Lambda(\boldsymbol{\theta}), \tag{A.3}$$

$$\lim_{|\theta| \to \infty} g(|\theta|, \boldsymbol{\theta}) = f(\boldsymbol{\theta}). \tag{A.4}$$

We will prove the following lemma.

LEMMA A.1.

$$\lim_{\Lambda \to Z^d} f_\Lambda(\boldsymbol{\theta}) = f(\boldsymbol{\theta}). \tag{A.5}$$

If $\Theta = \mathbb{R}$, then $\boldsymbol{\theta} \in \{-1, +1\}$. In this case, we will show that (A.5) implies the following uniform convergence.

LEMMA A.2. *If* $\Theta = \mathbb{R}$, *then*

$$\lim_{\Lambda \to Z^d} g_\Lambda(\theta) = g(\theta), \tag{A.6}$$

uniformly in θ, *for* θ *in the compactified real line.*

PROOF. It suffices to prove the lemma for $\boldsymbol{\theta} = +1$. That is, we will prove that the convergence in (A.6) is uniform in θ for θ in the compactified half-line $[0, +\infty]$. By Lemma A.1 and Proposition 3.1(iv), we have point-wise convergence in the compact set $[0, +\infty]$. We also have that $g(\theta)$ is continuous, and g_Λ, for each Λ, is monotone. These together with an easy extension of Dini's theorem [11], page 136, yield uniform convergence on the compactified half-line $[0, +\infty]$. □

Now, Lemma A.2 easily implies condition (2.21). Indeed, we have

$$\begin{aligned}\lim_{\Lambda \to Z^d} \sup_{|\theta| \ge A} |u_\Lambda(\theta)| &\le \sup_{|\theta| \ge A} |g(\theta) - g(A\boldsymbol{\theta})| \\ &= g(A\boldsymbol{\theta}) - f(\boldsymbol{\theta}).\end{aligned} \tag{A.7}$$

This and the continuity of g yield (2.21).

We do not know whether Lemma A.2 holds when Θ is not one-dimensional. But if it holds, then (A.7), and hence (2.21), also hold. Intuitively, a uniform convergence in (A.6) means that the finite-volume pressure $p_\Lambda(\theta)$ [or in general $p_{\Lambda, z}(\theta)$] have uniform asymptotes. It is for this reason that we feel that condition (2.21) is reasonable.

PROOF OF LEMMA A.1. The monotonicity of $g_\Lambda(|\theta|, \boldsymbol{\theta})$ in $|\theta|$ yields

$$\liminf_{\Lambda \to Z^d} f_\Lambda(\boldsymbol{\theta}) \le \limsup_{\Lambda \to Z^d} f_\Lambda(\boldsymbol{\theta}) \le f(\boldsymbol{\theta}). \tag{A.8}$$

By the variational principle, for any $\pi_\theta \in G_s(\theta)$ we have

$$\begin{aligned} f(\boldsymbol{\theta}) &\le g(|\theta|, \boldsymbol{\theta}) = |\theta| E_{\pi_\theta}\{\boldsymbol{\theta} \cdot A_U - m(\boldsymbol{\theta})\} + h(\pi_\theta) \\ &= |\theta| E_{\pi_\theta}\left\{ \lim_{\Lambda \to Z^d} \left[\boldsymbol{\theta} \cdot \frac{1}{|\Lambda|} U_\Lambda - m_\Lambda(\boldsymbol{\theta}) \right] \right\} + h(\pi_\theta) \\ &\le h(\pi_\theta). \end{aligned}$$

Let $\pi_{\boldsymbol{\theta}} \in G_a(\boldsymbol{\theta})$ and $\pi_{\theta_n} \in G_s(\theta_n)$ so that $\{\pi_{\theta_n}\}$ converges weakly to $\pi_{\boldsymbol{\theta}}$. Since $h(\cdot)$ is upper semicontinuous, we obtain

$$f(\boldsymbol{\theta}) \le h(\pi_{\boldsymbol{\theta}})$$

and therefore

$$f(\boldsymbol{\theta}) \le \inf\{h(\pi_{\boldsymbol{\theta}}) \colon \pi_{\boldsymbol{\theta}} \in G_a(\boldsymbol{\theta}), \text{stationary}\}. \tag{A.9}$$

By the finite-volume variational principle, we also have

$$g_\Lambda(|\theta|, \boldsymbol{\theta}) \ge |\theta| E_{\pi_{\boldsymbol{\theta}}}\left\{ \boldsymbol{\theta} \cdot \frac{1}{|\Lambda|} U_\Lambda - m_\Lambda(\boldsymbol{\theta}) \right\} + h_\Lambda\left(\pi_{\boldsymbol{\theta}}^{(\Lambda)}\right). \tag{A.10}$$

for any $\pi_{\boldsymbol{\theta}} \in G(\boldsymbol{\theta})$. We will show that, for sufficiently large Λ, the expectation in (A.10) is 0. Suppose not. Then we could find an $x(\Lambda)$ with $\pi_{\boldsymbol{\theta}}(x(\Lambda)) > 0$ and $\boldsymbol{\theta} \cdot (1/|\Lambda|) U_\Lambda(x(\Lambda)) < m_\Lambda(\boldsymbol{\theta})$. Since the interactions $\Phi^{(\alpha)}$ are of finite range, there exists, for large enough Λ, a subset $\Lambda_0 \subset \Lambda$ such that $\overline{\Lambda}_0 = \Lambda$, where the closure $\overline{\Lambda}_0$ is defined by

$$\overline{\Lambda}_0 = \{j \in Z^d: \exists\, A \subset Z^d: A \cap \Lambda \neq \varnothing,\, j \in A,\, \sup|\Phi(x(A))| \neq 0\}.$$

Then we would have $\pi_\theta(x(\Lambda - \Lambda_0)) > 0$ and $\pi_\theta(x(\Lambda_0)|x(\Lambda - \Lambda_0)) > 0$. These contradict the fact that $\pi_\theta(x(\Lambda_0)|x(\Lambda - \Lambda_0))$ concentrates on the set of configurations $x(\Lambda_0)$ maximizing $\theta \cdot \mathrm{U}_{\Lambda, x(\Lambda - \Lambda_0)}(x(\Lambda_0))$ when z is the free boundary condition. Therefore, we have

$$f_\Lambda(\theta) \geq h(\pi_\theta)$$

and hence

$$\liminf_{\Lambda \to Z^d} f_\Lambda(\theta) \geq \sup\{h(\pi_\theta): \pi_\theta \in G(\theta), \text{stationary}\}$$

$$\geq \sup\{h(\pi_\theta): \pi_\theta \in G_a(\theta), \text{stationary}\}.$$

Combining this with (A.8) and (A.9), we obtain the lemma. □

REFERENCES

[1] Ackley, D. H., Hinton, G. E. and Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Sci.* **9** 147–169.

[2] Almeida, M. and Gidas, B. (1989). A variational method for estimating parameters for MRF from complete or incomplete data. Preprint, Brown Univ.

[3] Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes. *Inequalities* **3** 1–8.

[4] Besag, J. (1977). Efficiency of pseudo-likelihood estimation for simple Gaussian fields. *Biometrika* **64** 616–618.

[5] Besag, J. (1986). On the statistical analysis of dirty pictures (with discussion). *J. Roy. Statist. Soc. Ser. B* **48** 259–302.

[6] Comets, F. (1986). Grandes déviations pour de champs de Gibbs sur Z^d. *C. R. Acad. Sci. Paris. Ser. I. Math.* **303** 511.

[7] Comets, F. (1989). Large deviation estimates for a conditional probability distribution. Applications to random interaction Gibbs measures. *Probab. Theory Related Fields* **80** 407–432.

[8] Comets, F. and Gidas, B. (1991). Asymptotics of maximum likelihood estimations for the Curie–Weiss model. *Ann. Statist.* **19** 557–578.

[9] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38.

[10] Derin, H. and Elliott, H. (1987). Modeling and segmentation of noisy and textured images using Gibbs random fields. *IEEE Trans. PAMI* **9** 39–55.

[11] Dieudonné, J. (1972). *Elements d'Analyse* **1**. Gauthiers-Villars, Paris.

[12] Dobrushin, R. L. (1972). Gibbs state describing co-existence of phases for a three-dimensional Ising model. *Theory Probab. Appl.* **17** 582–600.

[13] Föllmer, H. and Orey, S. (1988). Large deviations for the empirical field of a Gibbs measure. *Ann. Probab.* **16** 961–977.

[14] Frigessi, A. and Piccioni, M. (1990). Parameter estimation for two-dimensional Ising fields corrupted by noise. *Stochastic Process. Appl.* **34** 297–311.

[15] Geman, D. and Geman, S. (1988). *Maximum Entropy and Bayesian Methods in Science and Engineering.* (C. R. Smith and G. J. Erickson, eds.). Kluwer, Dordrecht.

[16] Geman, D., Geman, S., Graffigne, C. and Dong, P. (1990). Boundary detection by constraint optimization. *IEEE Trans. PAMI* **12**.

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

[18] GEMAN, S. and GRAFFIGNE, C. (1980). 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.

[19] GEMAN, S. and MCCLURE, D. E. (1985). Bayesian image analysis: An application to single photon emission tomography. In *Proc. Statist. Comput. Sect.* Amer. Statist. Assoc., Alexandria, Va.

[20] GEORGII, H. O. (1988). *Gibbs Measures and Phase Transitions*. de Gruyter, Berlin.

[21] GIDAS, B. (1988). Consistency of maximum likelihood and pseudo-likelihood estimators for Gibbs distribution. *Stochastic Differential Systems, Stochastic Control Theory, and Application* (W. Fleming and P. L. Lions, eds.). *IMA Vol. Math. Appl.* **10**. Springer, New York.

[22] GIDAS, B. (1989). A renormalization group approach to image processing problems. *IEEE Trans. PAMI* **11** 164–180.

[23] GIDAS, B. (1991). Parameter estimation for Gibbs distributions. I. Fully observed data. In *Markov Random Fields: Theory and Applications* (R. Chellapa and R. Jain, eds.). Academic, New York.

[24] GUYON, X. (1985). Estimation d'un champ par pseudo-vraisemblance conditionnelle: Etude asymptotique et application au cas Markovien. *Actes de la* 6^{eme} *Recontre Franco-Belge de Statisticiens, Bruxelles*.

[25] HINTON, G. E. and SEJNOWSKI, T. J. (1983). Optimal perceptual inference. *In Proc. IEEE Conf. Comput. Vision and Pattern Recognition*. IEEE Computer Society Press, Los Angeles.

[26] HOLSZTYNSKI, W. and SLAWNY, I. (1978). Peierls condition and number of ground states. *Comm. Math. Phys.* **61** 177–190.

[27] KÜNSCH, H. R. (1981). Thermodynamics and statistical analysis of Gaussian random fields. *Z. Wahrsch. Verw. Gebiete* **58** 407–421.

[28] LEBOWITZ, J. L. and PRESUTTI, E. (1976). Statistical mechanics of systems of unbounded spins. *Comm. Math. Phys.* **50** 195–218.

[29] LEDRAPPIER, F. (1977). Pressure and variational formula for random Ising model. *Comm. Math. Phys.* **56** 297.

[30] LEE, K. F. (1988). Large-vocabulary speaker-independent continuous speech recognition: The SPHINX system. Ph.D. dissertation, Dept. Computer Science, Carnegie Mellon Univ.

[31] LIPPMAN, A. (1986). A maximum entropy method for expert system construction. Ph.D. dissertation, Div. Applied Mathematics, Brown Univ.

[32] OLLA, S. (1988). Large deviation for Gibbs random fields. *Probab. Theory Related Fields* **77** 343–357.

[33] PICKARD, D. K. (1979). Asymptotic inference for Ising lattice. III. Non-zero fields and ferromagnetic states. *J. Appl. Probab.* **16** 12–24.

[34] PIROGOV, S. A. and SINAI, YA. G. (1976). Phase diagrams of classical lattice systems. *Teoret. Mat. Fiz.* **26** 61–76.

[35] POSSOLO, A. (1980). Estimation of binary Markov random fields. Preprint, Dept. Statistics, Univ. Washington.

[36] RUELLE, D. (1978). *Thermodynamic Formalism*. Addison-Wesley, Reading, Mass.

[37] SUNDBERG, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Statist.* **1** 49–58.

[38] YOUNÈS, L. (1988). Estimation and annealing for Gibbsian fields. *Ann. Inst. H. Poincaré Sect. B (N.S.)* **24** 269–294.

[39] YOUNÈS, L. (1988). Problemes d'estimation parametrique par des champs de Gibbs Markoviens. Application au traitement d'image. Thesis, Orsay, France.

[40] YOUNÈS, L. (1989). Parametric inference for imperfectly observed Gibbsian fields. *Probab. Theory Related Fields* **82** 625–645.

CENTRE DE MATHEMATIQUES APPLIQUÉES
ECOLE POLYTECHNIQUE
91128 PALAISEAU
FRANCE

DIVISION OF APPLIED MATHEMATICS
BROWN UNIVERSITY
PROVIDENCE, RHODE ISLAND 02912