The Annals of Applied Probability
1993, Vol. 3, No. 1, 103–136

A VARIATIONAL METHOD FOR ESTIMATING THE PARAMETERS OF MRF FROM COMPLETE OR INCOMPLETE DATA

BY MURILO P. ALMEIDA[1, 2] AND BASILIS GIDAS[1]

Brown University

We introduce a new method (to be referred to as the *variational method*, VM) for estimating the parameters of Gibbs distributions with random variables ("spins") taking values in a Euclidean space $\mathbb{R}^n$, $n \geq 1$, from complete or degraded data. The method applies also to the case of iid random variables governed by exponential families, and appears to be new even in this case. For complete data, the VM is computationally more efficient than, and as reliable as, the maximum pseudo-likelihood method. For incomplete data, the VM leads to an estimation procedure reminiscent of, but simpler than, the EM algorithm. In the former case, we show that under natural assumptions a certain form of the variational estimators is strongly consistent and asymptotically normal. We also present numerical experiments that demonstrate the computational efficiency and accuracy of the variational estimators.

1. Introduction. The massively computational tasks in image processing and computer vision problems [4, 11, 13, 16, 17], neural modelling and perceptual inference [1, 22] and speech recognition [27, 34] have created a need for more and more computationally efficient and reliable procedures for estimating the parameters of Gibbs [equivalently, Markov random fields (MRF)] and related distributions. From the theoretical point of view, these estimation problems generalize those of time-series analysis, and have given rise [6, 14, 20], to an interesting interplay between statistics and the phenomena of phase transitions in statistical mechanics.

The main methods that have been used for estimating the parameters of Gibbs distributions from *complete data* are: (a) maximum likelihood (ML) [12, 29, 41, 42, 20, 31, 35]; (b) maximum pseudo-likelihood (MPL) [3, 14, 18]; (c) a "coding" method [3]; and (d) a logistic-like method [9, 33]. The two main procedures that have been employed for estimating the parameters of Gibbs distributions from *incomplete* (noisy, degraded) *data* are: (i) maximum likelihood via the EM algorithm [8, 16, 39] and (ii) the method of moments [16]. From the theoretical point of view, consistency and asymptotic properties of various estimators have been studied in [7, 14, 18, 20]. From the computa-

Received November 1989; revised September 1991.
[1]Partially supported by ARO Contracts DAAL03-90-G-0033, DAAL03-86-K-0171 and ONR N00014-91-J-1021.
[2]Partially supported by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior-Brazil).
AMS 1991 *subject classifications*. 60K35, 60J99, 62H12, 62M40.
Key words and phrases. Super-stable Gibbs distributions, variational estimators, consistency, asymptotic normality.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Applied Probability*. www.jstor.org

tional point of view, the MPL procedure is the most efficient; it applies only to complete data, but to both continuous and discrete (categorical) random variables. For incomplete data, the ML method via the EM algorithm is computationally intensive, but feasible [8, 16].

In this paper we introduce a new computationally efficient method for estimating the parameters of Gibbs distributions with continuous random variables ("spins" [25, 36, 37, 32]) from complete or incomplete data. The method is also applicable to the case of iid random variables governed by exponential families, and appears to be new even in this case. For reasons to be justified later, we refer to the method as the *variational method* (VM) and to the corresponding estimators as *variational estimators* (VEs). In addition to the VM, we describe a class of Gibbs distributions (motivated from quantum field theory models [40]) that we used [2] successfully in synthesizing and representing textures, and appears to be appropriate for other spatial statistics applications.

We present two distinct, albeit conceptually related, versions of the VM. Both versions involve arbitrary vector statistics, and hence lead to a class of VEs. For complete data, the VEs are solutions of a system of linear equations. For a particular choice of the vector statistics, one of these systems [see (1.9)] has a structure similar to that of the Yule–Walker equations in time series analysis. For incomplete data, both versions lead to nonlinear equations, and to an estimation procedure reminiscent of the EM algorithm. We show (Section 3) that certain VEs of Gibbs distributions from complete data are strongly consistent regardless of the occurrence of phase transitions, and asymptotically normal under natural conditions. Our asymptotic normality result (Theorem 3.2) required the proof of a central limit theorem (Theorem 3.3) for dependent random variables; Theorem 3.3 appears to be of independent interest. Numerical experiments (Section 4) demonstrate the computational efficiency and accuracy of the VEs, and show that (for complete data) the VEs are less costly and as reliable as the MPL estimators.

The precise description of the VM for Gibbs distributions, with complete or incomplete data, is given in Section 2. In brief, the VM involves two basic steps. The first step uses the divergence theorem (equivalently, the integration by parts formula) to derive certain identities. The second step consists in replacing, in these identities, certain theoretical expectations by empirical expectations. In the present context, the divergence theorem has an interpretation in terms of a certain invariance of our probability laws (see later in this Introduction). The integration by parts formula was used previously in [5] (we thank one of the referees for bringing to our attention reference [5]) to derive moment recursion relations for a class of multimodal distributions; these relations are special cases of our equations. Our second step should be contrasted with the basic idea that underlies MPL, whereby one replaces "global" expectations with "local" expectations (using the Markov property) and evaluates the neighbors of a pixel at the observed data. In fact one form of the VEs may be derived from the MPL equations or generalizations thereof [see (2.19) and remarks following it].

The two basic steps of the VM are simpler and more transparent in the case of iid random variables governed by regular exponential families, and for this reason we present this case here: Let $\pi_\theta(dx)$ be a regular exponential family on $\mathbb{R}^n$ which is absolutely continuous with respect to the Lebesgue measure dx, with density $\pi_\theta(x)$ given by

$$\pi_\theta(x) = \frac{\exp\left(-\sum_{\alpha=1}^m \theta^{(\alpha)} U^{(\alpha)}(x)\right)}{Z(\theta)} \equiv \frac{\exp(-\theta \cdot U(x))}{Z(\theta)}, \tag{1.1}$$

where $U^{(\alpha)}(x)$, $\alpha = 1, \ldots, m$, are m sufficient statistics, $\theta = (\theta^{(1)}, \ldots, \theta^{(m)}) \in \mathbb{R}^m$ are the natural parameters (to be estimated from the data) with natural parameter space $\Theta \subseteq \mathbb{R}^m$, and $Z(\theta)$ is a normalizing constant. The first version of the VM is obtained by choosing m vector statistics $\mathbf{W}^{(\alpha)}(x) = (W_1^{(\alpha)}(x), \ldots, W_n^{(\alpha)}(x))$, $\alpha = 1, \ldots, m$, so that

$$\int_{\mathbb{R}^n} \nabla \cdot \left(\mathbf{W}^{(\alpha)}(x)\pi_\theta(x)\right) dx = 0, \qquad \alpha = 1, \ldots, m. \tag{1.2}$$

By the divergence theorem, (1.2) is an identity provided that the $\mathbf{W}^{(\alpha)}$'s are chosen so that the surface integral at infinity is 0 [see below for a "variational" interpretation of (1.2) in the present framework]. Writing (1.2) explicitly, we obtain the identities

$$\sum_{\beta=1}^m \theta^{(\beta)} \int_{\mathbb{R}^n} \{\mathbf{W}^{(\alpha)}(x) \cdot \nabla U^{(\beta)}(x)\}\pi_\theta(dx) = \int_{\mathbb{R}^n} \{\nabla \cdot \mathbf{W}^{(\alpha)}(x)\}\pi_\theta(dx), \tag{1.3}$$

$$\alpha = 1, \ldots, m.$$

If we replace the theoretical expectations in (1.3) by their empirical values, we obtain a linear system of equations (see Section 2) for the parameters $\theta^{(1)}, \ldots, \theta^{(m)}$. If the empirical estimate $\hat{T}^{(\alpha,\beta)}$ of the matrix

$$T^{(\alpha,\beta)} = \int_{\mathbb{R}^n} \{\mathbf{W}^{(\alpha)}(x) \cdot \nabla U^{(\beta)}(x)\}\pi_\theta(dx)$$

is invertible, then we obtain an estimator of $\theta = (\theta^{(1)}, \ldots, \theta^{(m)})$ that depends on the choice of the $\mathbf{W}^{(\alpha)}$'s. The invertibility (or lack thereof) depends on the $\mathbf{W}^{(\alpha)}$'s. In Section 3, we show that invertibility is guaranteed (even for Gibbs distributions) if we choose

$$\mathbf{W}^{(\alpha)}(x) = \nabla U^{(\alpha)}(x), \qquad \alpha = 1, \ldots, m. \tag{1.4}$$

If $\pi_\theta(dx)$ is Gaussian with (unknown) mean μ and (unknown) variance σ^2, $x \in \mathbb{R}$, the equations induced by (1.4) are exactly the ML equations. Furthermore, in Section 3 we show that the analogues of (1.6) and (1.4) for Gibbs distributions lead to VEs, which are strongly consistent and asymptotically normal.

Another choice of the $\mathbf{W}^{(\alpha)}$'s is: Choose m scalar statistics $F^{(\alpha)}(x)$, $\alpha = 1, \ldots, m$, and then define $\mathbf{W}^{(\alpha)}$'s so that

$$\nabla \cdot \mathbf{W}^{(\alpha)}(x) = F^{(\alpha)}(x), \qquad \alpha = 1, \ldots, m. \tag{1.5}$$

Then (1.3) reads

$$(1.6a)\quad \sum_{\beta=1}^{m} \theta^{(\beta)} \int_{\mathbb{R}^n} \{\mathbf{W}^{(\alpha)}(x) \cdot \nabla U^{(\beta)}(x)\} \pi_\theta(dx) = \int_{\mathbb{R}^n} F^{(\alpha)}(x) \pi_\theta(dx).$$

Again we replace the theoretical expectations with the empirical expectations, to obtain a linear system of equations for the $\theta^{(\alpha)}$'s. The choice (1.5) is natural, since the right-hand side of (1.6a) is related to the moment equations

$$(1.6b)\quad \int_{\mathbb{R}^n} f^{(\alpha)}(x) \pi_\theta(dx) = \text{empirical value of } F^{(\alpha)}, \qquad \alpha = 1, \ldots, m$$

and in particular to the ML equations corresponding to $F^{(\alpha)}(x) = U^{(\alpha)}(x)$, $\alpha = 1, \ldots, m$. In an obvious sense, the VE induced by (1.6a) is (if it exists) an approximate solution to the moment equations (1.6b). Numerical experiments (not reported in this paper) with Gibbs distributions, and with the choice $F^{(\alpha)}(x) = U^{(\alpha)}(x)$, $\alpha = 1, \ldots, m$, have given satisfactory results, but we have no theoretical results of consistency in this case (except in special situations).

The second version of the VM proceeds as follows: Choose m scalar statistics $G^{(\alpha)}(x)$, $\alpha = 1, \ldots, m$, and m vector statistics $\mathbf{W}^{(\alpha)}(x)$, $\alpha = 1, \ldots, m$, so that (1.2) with $\mathbf{W}^{(\alpha)}$ replaced by $\mathbf{W}^{(\alpha)}(x)G^{(\alpha)}(x)$ holds. The analogue of (1.3) now reads

$$(1.7)\quad \begin{aligned} &\sum_{\beta=1}^{m} \theta^{(\beta)} \left\langle [\mathbf{W}^{(\alpha)} \cdot \nabla U^{(\beta)}] G^{(\alpha)} \right\rangle^{(\theta)} \\ &\qquad = \left\langle [\nabla \cdot \mathbf{W}^{(\alpha)}] G^{(\alpha)} \right\rangle^{(\theta)} + \left\langle \mathbf{W}^{(\alpha)} \cdot \nabla G^{(\alpha)} \right\rangle^{(\theta)}, \end{aligned}$$

where $\langle \cdot \rangle^{(\theta)}$ denotes the expectation with respect to $\pi_\theta(dx)$. (Warning: Abusing notation—we use, throughout this paper, lower case letters to denote both the random variables and their realizations; in Section 2, we use upper case letters to denote a particular set of data.) Multiply both sides of (1.3) by $\langle G^{(\alpha)} \rangle^{(\theta)}$ and subtract the result from (1.7) to obtain the identity

$$(1.8)\quad \begin{aligned} &\sum_{\beta=1}^{m} \theta^{(\beta)} \operatorname{Cov}_\theta(\mathbf{W}^{(\alpha)} \cdot \nabla U^{(\beta)}, G^{(\alpha)}) \\ &\qquad = \left\langle \mathbf{W}^{(\alpha)} \cdot \nabla G^{(\alpha)} \right\rangle^{(\theta)} + \operatorname{Cov}_\theta(\nabla \cdot \mathbf{W}^{(\alpha)}, G^{(\alpha)}), \end{aligned}$$

$\alpha = 1, \ldots, m$, where $\operatorname{Cov}_\theta(\cdot, \cdot)$ denotes the covariance with respect to $\pi_\theta(dx)$. If the $\mathbf{W}^{(\alpha)}$'s are chosen so that $\mathbf{W}^{(\alpha)}(x) = x \in \mathbb{R}^n$ for all $\alpha = 1, \ldots, m$ then (1.8) becomes

$$(1.9)\quad \sum_{\beta=1}^{m} \theta^{(\beta)} \operatorname{Cov}_\theta(x \cdot \nabla U^{(\beta)}, G^{(\alpha)}) = \left\langle x \cdot \nabla G^{(\alpha)} \right\rangle^{(\theta)}, \qquad \alpha = 1, \ldots, m.$$

This identity has a structure which is similar to that of the Yule–Walker equations in time-series. If the theoretical expectations and covariances in (1.9) [or (1.8)] are replaced by their empirical values, we obtain a system of linear equations for $\theta^{(1)}, \ldots, \theta^{(m)}$. This gives another VE for θ. Identity (1.8)

with $G^{(\alpha)}(x) = U^{(\alpha)}(x)$, may be derived by differentiating (1.3) with respect to $\theta^{(\alpha)}$. Identities (1.6) and (1.9) easily yield the moment recursion relations of [5].

In the present framework, identity (1.2) has the following interpretation which justifies the term "variational method": The measure $\pi_\theta(dx)$ is invariant under the transformation ("variation" of random variables) $x_i \to x_i + \varepsilon_i$, $\varepsilon_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$, but its sufficient statistics $U^{(\alpha)}(x)$ are not. Thus, if $\phi(x) = \phi(x_1, \ldots, x_n)$ is a scalar statistic, then its expectation

$$\langle \phi \rangle^{(\theta)} = \int_{\mathbb{R}^n} \phi(x_1, \ldots, x_n) \pi_\theta(x_1, \ldots, x_n)\, dx$$

is invariant under the particular transformation $x_i \to x_i + \varepsilon$, $x_j \to x_j$ for all $j \neq i$, $\varepsilon \in \mathbb{R}$, that is,

$$\langle \phi \rangle^{(\theta)} = \int_{\mathbb{R}^n} \phi(x_1, \ldots, x_i + \varepsilon, \ldots, x_n) \pi_\theta(x_1, \ldots, x_i + \varepsilon, \ldots, x_n)\, dx. \tag{1.10}$$

Differentiating both sides of (1.10) with respect to ε and setting $\varepsilon = 0$, we obtain (formally)

$$\int_{\mathbb{R}^n} \frac{\partial}{\partial x_i} (\phi(x) \pi_\theta(x))\, dx = 0. \tag{1.11}$$

Applying this with $\phi(x) = W_i^{(\alpha)}(x)$, $i = 1, 2, \ldots, n$, and summing over i, we obtain (1.2). The above differentiation (with respect to ε) is formally equivalent to the standard procedure for deriving the "variational" (Euler–Lagrange) equations in partial differential equations. Here, this procedure amounts in expanding (formally) the integrand in (1.10) in a Taylor series in ε, and then setting to zero the linear in ε term to obtain (1.11).

The preceding arguments show that the first version of the VM has its roots in the "variation" of the variables. The second version stems from the "variation" of the variable and of the parameters [recall that (1.8) with $G^{(\alpha)} = U^{(\alpha)}$ is obtained by differentiating (1.3) with respect to $\theta^{(\alpha)}$].

The VM as previously presented is appropriate for continuous random variables taking values in $\mathbb{R}^n$ or in a compact manifold without boundaries (so that no boundary terms appear in the divergence theorem). If the manifold has boundaries (e.g., a bounded interval in $\mathbb{R}$), then the procedure still applies but the equations may not be simple because of the boundary terms. The VM can be formulated also for categorical variables (e.g., binary Ising model). But in this case, the equations are nonlinear even for complete data; for the Ising model the equations have a structure similar to those of MPL, but for general models the equations are not very convenient for estimation.

The VM is ideally suitable for Gibbs distributions with superstable regular interactions [25, 36, 37], to be briefly described in Section 2. Here we present a particular class of these distributions which has been successfully used [2] in certain image processing tasks, and has recently been recommended [24] as priors in emission tomography: Let $J = \{J(i-j)\}_{i,j \in \mathcal{Y}^d}$ be a positive definite

matrix such that

$$J(j) = J(-j) \quad \text{for all } j \in \mathcal{Z}^d, \qquad \sum_{j \in \mathcal{Z}^d} |J(j)| < \infty.$$

In the experiments of Section 4, $d = 2$ and $J(j) = 0$ for $|j|$ larger than some R_0. Let

$$p(x;\lambda) = \lambda^{(2m)}x^{2m} + \lambda^{(2m-1)}x^{2m-1} + \cdots + \lambda^{(1)}x, \qquad x \in \mathbb{R} \tag{1.12}$$

be a polynomial of even degree with $\lambda^{(2m)} > 0$. With each pixel $i \in \mathcal{Z}^d$, we associate a random variable $x_i \in \mathbb{R}$. Let Λ be a window (typically a hypercube) in $\mathcal{Z}^d$. Consider the energy function (Hamiltonian) in the finite window Λ:

$$\begin{aligned} H_\Lambda^{(\theta)}(x(\Lambda)) &= \tfrac{1}{2} \sum_{i,j \in \Lambda} J(i-j)x_i x_j + \sum_{i \in \Lambda} p(x_i;\lambda), \\ x(\Lambda) &= \{x_i : i \in \Lambda\}, \\ \theta &= (J, \lambda). \end{aligned} \tag{1.13}$$

The distribution of $x(\Lambda) = \{x_i : i \in \Lambda\}$ is given by the "finite window" Gibbs distribution

$$\pi_\theta(dx(\Lambda)) = \frac{\exp\left[-H_\Lambda^{(\theta)}(x(\Lambda))\right]}{Z_\Lambda(\theta)} \prod_{i \in \Lambda} dx_i. \tag{1.14}$$

For fixed Λ, the distribution (1.14) is of the exponential family, but the random variables $\{x_i : i \in \Lambda\}$ are dependent. In fact, for some values of the parameter $\theta = (J, \lambda)$, these models exhibit the phenomena of phase transitions (and long range dependence) as $\Lambda \to \mathcal{Z}^d$. It is these properties that make this class of distributions suitable for spatial statistics applications such as image processing tasks and Bayesian regularization [2]. A particular subclass of (1.14) behaves qualitatively and quantitatively like binary Ising models [see (4.2) and remarks following (4.2)]. For $J(j) \equiv 0$, $j \in \mathcal{Z}^d$, the random variables are iid whose density coincides with those of type **N** and **G** in [5].

The general class of Gibbs distributions described in Section 2 is (for finite Λ) of the form (1.14) with $H_\Lambda^{(\theta)}$'s linear in $\theta \in \mathbb{R}^m$. We will be interested in estimating the parameters θ from a single realization (complete data case) $X(\Lambda) = \{X_i : i \in \Lambda\}$, or a single partial observation $Y(\Lambda) = \{Y_i : i \in \Lambda\}$. Consistency and asymptotic normality are studied as $\Lambda \to \mathcal{Z}^d$; hence, this study generalizes that of time series analysis.

Since the VEs of Gibbs distributions from complete data are solutions of linear equations, they are computationally more efficient than the MPL estimators (see Section 4 for comparison of CPU times). Furthermore, our numerical experiments show that the VEs are as accurate as the MPL estimators. The comparison of the asymptotic (as $\Lambda \to \mathcal{Z}^d$) variances of the VEs and MPL estimators (when both are asymptotically normal), that is, the estimation of their asymptotic relative efficiency, is an interesting open problem; this is similar to the analogous (and still open) problem for ML and MPL estimators. These problems appear to be subtle even for the models (1.11) and their

simpler versions (4.2). They lead to certain correlation or moment inequalities whose proof (if true) seems to be delicate.

The organization of this paper is as follows: In Section 2 we set up our notation and describe the variational method for Gibbs distributions. Section 3 contains the theorems and proofs of consistency and asymptotic normality for certain VEs. Numerical experiments with the VEs and comparison with MPL estimators are reported in Section 4.

2. The variational method for parameter estimation. In this section we describe in detail the variational method (VM) for estimating the parameters of Gibbs distributions with unbounded continuous random variables from complete and incomplete data. For the case of complete data we also derive a generalization of the maximum pseudo-likelihood (MPL) equations. But first we introduce some basic definitions and properties for superstable regular Gibbs distributions [25, 26, 36, 37, 23].

A. *Superstable regular Gibbs distributions.* We consider the configurations (or state) spaces

$$\Omega = (\mathbb{R}^n)^{\mathscr{Z}^d}, \qquad \Omega_\Lambda = (\mathbb{R}^n)^\Lambda,$$

where Λ is a finite window ("volume") in $\mathscr{Z}^d$. The Gibbs distributions treated in this paper are defined on Ω. They are limits (in a weak sense) of the so-called finite-volume Gibbs distributions defined on Ω_Λ. Configurations (states) in Ω and Ω_Λ will be denoted by $x = \{x_i\colon\ i \in \mathscr{Z}^d\}$ and $x(\Lambda) = \{x_i\colon\ i \in \Lambda\}$, respectively. For Gibbs distributions in general, $\mathbb{R}^n$ is replaced by an arbitrary state space Ω_0. We assume that a reference Borel measure μ_0 on $\mathbb{R}^n$ is given such that

$$\int e^{-\alpha|x|^2}\, d\mu_0(x) < +\infty$$

for all $\alpha > 0$. We set $d\mu_{0,\Lambda}(x(\Lambda)) = \prod_{i\in\Lambda} d\mu_0(x_i)$.

Gibbs distributions are defined in terms of potentials (interactions). Let $\mathscr{V}$ denote the set of all finite subsets of $\mathscr{Z}^d$. A potential is a map $\Phi\colon \bigcup_{V\in\mathscr{V}} \Omega_V \to \mathbb{R}$ so that $\Phi(x(V))$ is $\mu_{0,V}$-measurable. We will always assume that the potentials are invariant (stationary) under translations of $\mathscr{Z}^d$.

The energy associated with a configuration x in a finite-volume $\Lambda \in \mathscr{V}$ is defined by

$$U_\Lambda(x) \equiv U_\Lambda(x(\Lambda)) = \sum_{V\subset\Lambda} \Phi(x(\Lambda)). \tag{2.1}$$

The potential Φ (or the energy) is said to be superstable [36, 37], if there exist $A > 0$, $B \in \mathbb{R}$ such that

$$U_\Lambda(x) \geq \sum_{i\in\Lambda} \left(A|x_i|^2 + B\right) \tag{2.2}$$

for all $\Lambda \in \mathscr{V}$. For $\Lambda, \Lambda' \in \mathscr{V}$, $\Lambda \cap \Lambda' = \varnothing$, the interaction between Λ and Λ' is

defined by

$$(2.3)\quad I_{\Lambda,\Lambda'}(x) = U_{\Lambda\cup\Lambda'}(x) - U_{\Lambda}(x) - U_{\Lambda'}(x) = \sum_{\substack{V\subset\Lambda\cup\Lambda' \\ V\cap\Lambda\neq\varnothing\neq V\cap\Lambda'}} \Phi(x(V)).$$

The potential Φ (or the energy U) is said to be regular [36, 37], if there exists a monotonically decreasing function $\Psi\colon \mathscr{Z}_+ \to \mathbb{R}_+$, with

$$\sum_{j\in\mathscr{Z}^d} \Psi(|j|) < +\infty,$$

such that

$$(2.4)\quad |I_{\Lambda,\Lambda'}(x)| \leq \tfrac{1}{2} \sum_{i\in\Lambda} \sum_{j\in\Lambda'} \Psi(|i-j|)\left(|x_i|^2 + |x_j|^2\right)$$

for all $\Lambda, \Lambda' \in \mathscr{V}$, $\Lambda \cap \Lambda' = \varnothing$. Here $|\cdot|$ denotes a norm on $\mathscr{Z}^d$ defined by $|j| = \max\{|j_1|, \ldots, |j_d|\}$.

The energy $U_{\Lambda}(x(\Lambda))$, $\Lambda \in \mathscr{V}$, is called the finite-volume energy with *free boundary conditions* (*b.c.*) on Λ^c. If Λ is a torus, then $U_{\Lambda}(x(\Lambda))$ is the energy with periodic b.c. We now introduce other b.c. [25]. A configuration $z \in \Omega$ is said to be an admissible b.c. if

$$\begin{aligned} U_{\Lambda,z}(x(\Lambda)) &\equiv U_{\Lambda}(x(\Lambda)) + I_{\Lambda,\Lambda^c}(x(\Lambda) \vee z(\Lambda^c)) \\ &= \sum_{\substack{V\in\mathscr{V} \\ V\cap\Lambda\neq\varnothing}} \Phi(x(V) \vee z(V)) \end{aligned} \quad (2.5)$$

is well defined and

$$Z_{\Lambda,z}(\Phi) = \int_{\Omega_\Lambda} \exp[-U_{\Lambda,z}(x(\Lambda))]\, d\mu_{0,\Lambda}(x(\Lambda)) < +\infty.$$

The configuration $x(V) \vee z(V)$ in (2.5) is defined by

$$(2.6)\quad (x(V) \vee z(V))_i = \begin{cases} x_i, & \text{if } i \in V \cap \Lambda, \\ z_i, & \text{if } i \in V \cap \Lambda^c. \end{cases}$$

Examples of admissible b.c. are given in [25] (Section 3). The case $z = 0$ will be referred to as Dirichlet b.c.

The finite-volume Gibbs distribution in $\Lambda \in \mathscr{V}$ with (admissible) b.c. z is defined by

$$(2.7a)\quad \pi_{\Phi,\Lambda,z}(dx(\Lambda)) = \pi_{\Phi,\Lambda,z}(x(\Lambda))\, d\mu_{0,\Lambda}(x(\Lambda)),$$

$$(2.7b)\quad \pi_{\Phi,\Lambda,z}(x(\Lambda)) = \frac{\exp[-U_{\Lambda,z}(x(\Lambda))]}{Z_{\Lambda,z}(\Phi)}.$$

A probability measure π on Ω is called a *Gibbs distribution* relative to the potential Φ, if the conditional distribution $\pi(dx(\Lambda)|z(\Lambda^c))$ for any $\Lambda \in \mathscr{V}$ and admissible b.c. z has a version given by $\pi(dx(\Lambda)|z(\Lambda^c)) = \pi_{\Phi,\Lambda,z}(dx(\Lambda))$. It is

called *tempered* if

$$\pi\left\{ \bigcup_N \bigcap_n \left\{ \sum_{|i|\le n} |x_i|^2 \le N^2(2n+1)^d \right\} \right\} = 1.$$

The set of all Gibbs distributions associated with Φ will be denoted by $G(\Phi)$, and its subset composed of all stationary distributions will be denoted by $G_0(\Phi)$. Lebowitz and Presutti have shown [25] that if Φ is superstable and regular, then $G_0(\Phi)$ is nonempty. Furthermore, the elements of $G_0(\Phi)$ are *regular* in the sense that

$$\frac{d\pi^{(\Lambda)}(x(\Lambda))}{d\mu_{0,\Lambda}} \le \exp\left[-\sum_{i\in\Lambda} \left(\gamma|x_i|^2 - \delta \right) \right] \tag{2.8}$$

for some $\gamma > 0$, $\delta \in \mathbb{R}$. Here $d\pi^{(\Lambda)}(x(\Lambda))$ denotes the restriction for π to Ω_Λ. If $G_0(\Phi)$ is not a singleton, then we say that a phase transition occurs. The set $G_0(\Phi)$ is convex and its extremal points $\mathscr{E}_0(\Phi)$ are the ergodic Gibbs distributions. Föllmer has shown [10] that any $\pi \in G_0(\Phi)$ has an *ergodic decomposition*, that is,

$$\pi(\cdot) = \int_{\mathscr{E}_0(\Phi)} P^{(\xi)}(\cdot)\, d\rho_\pi(\xi), \tag{2.9}$$

where $d\rho_\pi(\cdot)$ is a probability measure on $\mathscr{E}_0(\Phi)$.

In this paper we fix m superstable regular potentials $\Phi^{(\alpha)}$, $\alpha = 1, \dots, m$, and consider Gibbs distributions parametrized by a vector $\theta = (\theta^{(1)}, \dots, \theta^{(m)}) \in \mathbb{R}^m$. The parameter space Θ is either $\mathbb{R}^m$ or a convex subset of $\mathbb{R}^m$. Let $U_{\Lambda,z}^{(\alpha)}$ be the energy, (2.5), associated with $\Phi^{(\alpha)}$ and with b.c. z. The finite-volume Gibbs distributions parametrized by $\theta \in \Theta$ are obtained from (2.7) by replacing $U_{\Lambda,z}$ by

$$H_{\Lambda,z}^{(\theta)}(x(\Lambda)) = \theta \cdot U_{\Lambda,z}(x(\Lambda)) = \sum_{\alpha=1}^{m} \theta^{(\alpha)} U_{\Lambda,z}^{(\alpha)}(x(\Lambda)), \tag{2.10}$$

that is,

$$\pi_{\theta,\Lambda,z}(x(\Lambda)) = \frac{\exp\left[-H_{\Lambda,z}^{(\theta)}(x(\Lambda)) \right]}{Z_{\Lambda,z}(\theta)} = \frac{\exp\left[-\theta \cdot U_{\Lambda,z}(x(\Lambda)) \right]}{Z_{\Lambda,z}(\theta)}. \tag{2.11}$$

$H_{\Lambda,z}^{(\theta)}$ will be referred to as the finite-volume Hamiltonian with b.c. z. For $z = 0$ or periodic b.c., we will suppress the index z. The set of all Gibbs distributions corresponding to a particular value of θ will be denoted by $G(\theta)$, and the set of all stationary (with respect to translation on $\mathscr{Z}^d$) Gibbs distributions will be denoted by $G_0(\theta)$.

B. *The variational method for complete data.* Here and in the remainder of the paper we will assume that the reference measure $d\mu_0$ on $\mathbb{R}^n$ is the Lebesgue measure, that is, $d\mu_0(x_i) = dx_i$. We will also assume that the potentials $\Phi^{(\alpha)}$, $\alpha = 1, \dots, m$, are of finite range, that is, $\Phi(x(V)) = 0$ when-

ever the diameter, $d(V) = \max\{|i-j|: i, j \in V\}$, of V is larger than a constant (integer) R_0, called here radius of interactions (most of our arguments can be extended to infinite range interactions). We will denote by $\mathcal{N}_i$ the set of pixels $j \in \mathcal{L}^d$, $j \neq i$, that interact with i (note that if $j \in \mathcal{N}_i$, then $|i-j| \leq R_0$). Let $\Lambda \subset \mathcal{L}^d$ be a finite window (volume) with complement $\Lambda^c = \mathcal{L}^d \setminus \Lambda$. The interior Λ^0 of Λ is defined by $\Lambda^0 = \{j \in \Lambda: \mathcal{N}_j \cap \Lambda^c = \varnothing\}$ and the (exterior) boundary $\partial\Lambda$ of Λ by $\partial\Lambda = \{j \in \Lambda^c: \mathcal{N}_j \cap \Lambda \neq \varnothing\}$. For $i \in \Lambda$ we define $\mathcal{U}_{i,\Lambda}^{(\alpha)}$ (in shorthand notation $\mathcal{U}_i^{(\alpha)}$) by

$$\text{(2.12a)} \quad \mathcal{U}_i^{(\alpha)} \equiv \mathcal{U}_{i,\Lambda}^{(\alpha)}(x_i, x(\mathcal{N}_i) \vee z(\mathcal{N}_i)) = \sum_{i \in V} \Phi^{(\alpha)}(x(V) \vee z(V)) \qquad i \in \Lambda,\ \alpha = 1, \ldots, m,$$

where the configurations $x(V) \vee z(V)$ and $x(\mathcal{N}_i) \vee z(\mathcal{N}_i)$ are defined relative to Λ by (2.6). Note that

$$\text{(2.13)} \qquad x(\mathcal{N}_i) \vee z(\mathcal{N}_i) = x(\mathcal{N}_i) \quad \text{if } i \in \Lambda^0.$$

Hence

$$\text{(2.12b)} \qquad \mathcal{U}_i^{(\alpha)} = \mathcal{U}_i^{(\alpha)}(x_i, x(\mathcal{N}_i)) \quad \text{if } i \in \Lambda^0,$$

is independent of z (and Λ) if $i \in \Lambda^0$. We also define

$$\text{(2.14)} \quad H_i^{(\theta)} \equiv H_{i,\Lambda}^{(\theta)}(x_i, x(\mathcal{N}_i) \vee z(\mathcal{N}_i)) = \sum_{\alpha=1}^{m} \theta^{(\alpha)} \mathcal{U}_i^{(\alpha)}, \qquad i \in \Lambda.$$

The local characteristics of $\pi_{\theta,\Lambda,z}(dx(\Lambda))$ are given by

$$\text{(2.15a)} \quad d\pi_{\theta,i} \equiv \pi_{\theta,\Lambda}(dx_i | x(\mathcal{N}_i) \vee z(\mathcal{N}_i)) = \frac{\exp[-H_i^{(\theta)}]}{\int_{\mathbb{R}^n} \exp[-H_i^{(\theta)}]}, \qquad i \in \Lambda.$$

Because of (2.13) we have for $i \in \Lambda^0$,

$$\text{(2.15b)} \quad d\pi_{\theta,i} = \pi_\theta(dx_i | x(\mathcal{N}_i)) = \frac{\exp[-H_i^{(\theta)}(x_i, x(\mathcal{N}_i))]}{\int_{\mathbb{R}^n} \exp[-H_i^{(\theta)}(\xi_i, x(\mathcal{N}_i))]\, d\xi_i}\, dx_i.$$

These are also the local characteristics of any $\pi_\theta \in G(\theta)$.

Throughout this and the next subsection the finite window (volume) Λ and the boundary condition z are fixed. We are interested in estimating the parameter vector $\theta = (\theta^{(1)}, \ldots, \theta^{(m)})$ from a single realization $X(\Lambda) = \{X_i: i \in \Lambda\}$. The variational method (VM) proceeds as in the case of (1.1): We choose m vector statistics $\mathbf{W}_\Lambda^{(\alpha)}(x(\Lambda)) = \{W_i^{(\alpha)}\}_{i \in \Lambda}$ (possibly z dependent) whose components are "localized," that is, they have the form

$$\text{(2.16)} \quad W_i^{(\alpha)} = W_{i,\Lambda}^{(\alpha)}(x_i, x(\mathcal{N}_i) \vee z(\mathcal{N}_i)), \qquad i \in \Lambda,\ \alpha = 1, \ldots, m,$$

where $x(\mathcal{N}_i) \vee z(\mathcal{N}_i)$ satisfies (2.13). We assume [compare with (1.2)]

$$\text{(2.17)} \quad \int_{\Omega_\Lambda} \{\nabla \cdot (\mathbf{W}_\Lambda^{(\alpha)}(x(\Lambda)) \pi_{\theta,\Lambda,z}(x(\Lambda)))\}\, dx(\Lambda) = 0, \qquad \alpha = 1, \ldots, m.$$

As in (1.3), this gives the identities

$$\sum_{\beta=1}^{m} \theta^{(\beta)} \left\{ \sum_{i\in\Lambda} \int_{\Omega_\Lambda} W_i^{(\alpha)} \frac{\partial \mathscr{U}_i^{(\beta)}}{\partial x_i} \pi_{\theta,\Lambda,z}(dx(\Lambda)) \right\}$$

$$= \sum_{i\in\Lambda} \int_{\Omega_\Lambda} \frac{\partial W_i^{(\alpha)}}{\partial x_i} \pi_{\theta,\Lambda,z}(dx(\Lambda)), \qquad \alpha = 1,\ldots,m. \tag{2.18}$$

We will use this identity to derive the two versions of the VM. But first we use (2.18) to derive a system of (nonlinear) equations that generalize the MPL equations: Using the local characteristics (2.15) we write the left-hand side of (2.18) in terms of an "inner" (local) expectation with respect to $d\pi_{\theta,i}$, and an "outer" (global) expectation with respect to $\pi_{\theta,\Lambda-\{i\},z}(dx(\Lambda - \{i\}))$. Then we replace the outer expectation with its empirical value, and on the right-hand side of (2.18) we replace the expectation with its empirical value. This leads to the equations

$$\sum_{\beta=1}^{m} \theta^{(\beta)} \left\{ \sum_{i\in\Lambda} \int_{\mathbb{R}^n} W_{i,\Lambda}^{(\alpha)}(x_i, X(\mathscr{N}_i) \vee z(\mathscr{N}_i)) \right.$$

$$\left. \times \frac{\partial \mathscr{U}_{i,\Lambda}^{(\beta)}(x_i, X(\mathscr{N}_i) \vee z(\mathscr{N}_i))}{\partial x_i} \pi_{\theta,\Lambda}(dx_i | X(\mathscr{N}_i) \vee z(\mathscr{N}_i)) \right\} \tag{2.19}$$

$$= \sum_{i\in\Lambda} \frac{\partial}{\partial x_i} W_{i,\Lambda}^{(\alpha)}(X_i, X(\mathscr{N}_i) \vee z(\mathscr{N}_i)), \qquad \alpha = 1,\ldots,m.$$

[Recall that $\mathscr{U}_{i,\Lambda}^{(\alpha)}$ and $X(\mathscr{N}_i) \vee z(\mathscr{N}_i)$ satisfy (2.12b) and (2.13), respectively.] These nonlinear equations generalize the MPL equations. Indeed, if the $\mathbf{W}_\Lambda^{(\alpha)}$ are chosen so that

$$\frac{\partial W_i^{(\alpha)}}{\partial x_i} = \mathscr{U}_i^{(\alpha)}, \qquad i \in \Lambda, \alpha = 1,\ldots,m, \tag{2.20}$$

then it can easily be shown that (2.19) are equivalent to the MPL equations. The first version of the VM goes a step further and replaces the local expectations in (2.19) by their empirical values. This amounts to replacing the theoretical expectations in (2.18) by the empirical expectations. This gives the system of linear equations

$$\sum_{\beta=1}^{m} \theta^{(\beta)} \left\{ \sum_{i\in\Lambda} W_{i,\Lambda}^{(\alpha)}(x_i, X(\mathscr{N}_i) \vee z(\mathscr{N}_i)) \frac{\partial \mathscr{U}_{i,\Lambda}^{(\beta)}(x_i, X(\mathscr{N}_i) \vee z(\mathscr{N}_i))}{\partial x_i} \right\}$$

$$= \sum_{i\in\Lambda} \frac{\partial}{\partial x_i} W_{i,\Lambda}^{(\alpha)}(x_i, X(\mathscr{N}_i) \vee z(\mathscr{N}_i)), \qquad \alpha = 1,\ldots,m. \tag{2.21}$$

If the matrix

$$\hat{T}^{(\alpha,\beta)}_{\Lambda,z}(X(\Lambda)) = \sum_{i\in\Lambda} W_i^{(\alpha)} \frac{\partial \mathscr{U}_i^{(\beta)}}{\partial x_i} \tag{2.22}$$

is invertible, then (2.21) gives the first VE, $\hat{\theta}_\Lambda$, of θ.

A natural choice for the $\mathbf{W}_\Lambda^{(\alpha)}$'s is as in (1.4), that is,

$$\mathbf{W}_\Lambda^{(\alpha)}(x(\Lambda)) = \nabla U^{(\alpha)}_{\Lambda,z}(x(\Lambda)), \qquad \alpha = 1,\dots,m \tag{2.23a}$$

or in terms of components,

$$W_i^{(\alpha)} = \frac{\partial \mathscr{U}_i^{(\alpha)}}{\partial x_i}. \tag{2.23b}$$

In this case, (2.21) reads

$$\sum_{\beta=1}^{m} \theta^{(\beta)} \left\{ \sum_{i\in\Lambda} \frac{\partial \mathscr{U}_i^{(\alpha)}}{\partial x_i} \frac{\partial \mathscr{U}_i^{(\beta)}}{\partial x_i} \right\} = \sum_{i\in\Lambda} \frac{\partial^2 \mathscr{U}_i^{(\alpha)}}{\partial x_i^2}, \qquad \alpha = 1,\dots,m, \tag{2.24}$$

where we have used the notation of (2.12a) [the arguments in (2.24) are evaluated at the data $X(\Lambda)$]. In Section 3 we show that for Λ sufficiently large the solution of (2.24) exists and, under suitable conditions, is strongly consistent and asymptotically normal as $\Lambda \to \mathscr{Z}^d$. In Section 4 we report numerical results that demonstrate the computational efficiency and accuracy of these estimators.

Another choice of the $\mathbf{W}_\Lambda^{(\alpha)}$'s is a generalization of (2.20): Let $F_\Lambda^{(\alpha)}(x(\Lambda))$ be a (possibly z dependent) scalar statistic built out of local pieces, that is,

$$F_\Lambda^{(\alpha)}(x(\Lambda)) = \sum_{i\in\Lambda} F_{i,\Lambda}^{(\alpha)}(x_i, x(\mathscr{N}_i) \vee z(\mathscr{N}_i)), \qquad \alpha = 1,\dots,m. \tag{2.25}$$

[$F_\Lambda^{(\alpha)}(x(\Lambda))$ may be chosen to be independent of z, in which case, $x(\mathscr{N}_i)$ in (2.25) is replaced by $x(\mathscr{N}_i \cap \Lambda)$.] Now we choose the $W_\Lambda^{(\alpha)}$'s so that

$$\frac{\partial W_{i,\Lambda}^{(\alpha)}}{\partial x_i} = F_{i,\Lambda}^{(\alpha)}, \qquad i \in \Lambda \ \alpha = 1,\dots,m. \tag{2.26}$$

Note that in this case the right-hand side of (2.18) reads

$$\int_{\Omega_\Lambda} \nabla \cdot W_\Lambda^{(\alpha)}(x(\Lambda)) \pi_{\theta,\Lambda,z}(dx(\Lambda)) = \int_{\Omega_\Lambda} F_\Lambda^{(\alpha)}(x(\Lambda)) \pi_{\theta,\Lambda,z}(dx(\Lambda)).$$

Hence (2.21) gives an approximate solution of the moment equation

$$\int_{\Omega_\Lambda} F_\Lambda^{(\alpha)}(x(\Lambda)) \pi_{\theta,\Lambda,z}(dx(\Lambda)) = F_\Lambda^{(\alpha)}(X(\Lambda)), \qquad \alpha = 1,\dots,m. \tag{2.27}$$

If $F_\Lambda^{(\alpha)}(x(\Lambda)) = \mathscr{U}_{\Lambda,z}^{(\alpha)}(x(\Lambda))$, $\alpha = 1,\dots,m$, (2.27) are exactly the ML equations.

The second version of the VM proceeds as in the derivation of (1.8) [and (1.9)]: We choose m vector statistics $\mathbf{W}_\Lambda^{(\alpha)}$ localized as in (2.16), and m scalar statistics $G_\Lambda^{(\alpha)}$ built of local pieces as in (2.25). Then proceeding as in the

derivation of (1.8), we arrive at the identities ($\alpha = 1, \ldots, m$)

$$\sum_{\beta=1}^{m} \theta^{(\beta)} \operatorname{Cov}_{\theta,\Lambda,z}\left(\mathbf{W}_\Lambda^{(\alpha)} \cdot \nabla U_{\Lambda,z}^{(\beta)}, G_\Lambda^{(\alpha)}\right)$$

$$= \left\langle \mathbf{W}_\Lambda^{(\alpha)} \cdot \nabla G_\Lambda^{(\alpha)} \right\rangle_{\Lambda,z}^{(\theta)} + \operatorname{Cov}_{\theta,\Lambda,z}\left(\nabla \cdot \mathbf{W}_\Lambda^{(\alpha)}, G_\Lambda^{(\alpha)}\right), \tag{2.28}$$

where $\operatorname{Cov}_{\theta,\Lambda,z}(\cdot,\cdot)$ and $\langle \cdot \rangle_{\Lambda,z}^{(\theta)}$ denote covariance and expectation, respectively, with respect to $\pi_{\theta,\Lambda,z}(dx(\Lambda))$. If $\mathbf{W}_\Lambda^{(\alpha)}(x(\Lambda)) = x(\Lambda)$ for $\alpha = 1, \ldots, m$, then (2.28) becomes

$$\sum_{\beta=1}^{m} \theta^{(\beta)} \left\{ \sum_{i,j\in\Lambda} \operatorname{Cov}_{\theta,\Lambda,z}\left(x_i \frac{\partial \mathscr{U}_i^{(\beta)}}{\partial x_i}, G_j^{(\alpha)} \right) \right\} = \sum_{i\in\Lambda} \left\langle x_i \frac{\partial G_i^{(\alpha)}}{\partial x_i} \right\rangle_{\Lambda,z}^{(\theta)} \tag{2.29}$$

$$\alpha = 1, \ldots, m.$$

As in the iid case (1.9), the structure of the identities (2.29) is similar to that of the Yule–Walker equations in time-series. A VE is obtained if we replace the theoretical covariance and expectation in (2.29) by their empirical values. For the iid case (1.9), these empirical values are straightforward and lead to the following equations: If $X^{(1)}, \ldots, X^{(N)}$ ($X^{(i)} \in \mathbb{R}^n$) are the iid data, then

$$\sum_{\beta=1}^{m} \theta^{(\beta)} \frac{1}{N} \sum_{j=1}^{N} \left\{ \left[X^{(j)} \cdot \nabla U^{(\beta)}(X^{(j)}) - \bar{A}_N^{(\beta)} \right] \left[G^{(\alpha)}(X^{(j)}) - \bar{G}_N^{(\alpha)} \right] \right\}$$

$$= \frac{1}{N} \sum_{j=1}^{N} X^{(j)} \cdot \nabla G^{(\alpha)}(x^{(j)}), \qquad \alpha = 1, \ldots, m, \tag{2.30}$$

where

$$\bar{A}_N^{(\beta)} = \frac{1}{N} \sum_{j=1}^{N} X^{(j)} \cdot \nabla U^{(\beta)}(X^{(j)}), \qquad \bar{G}_N^{(\alpha)} = \frac{1}{N} \sum_{j=1}^{N} G^{(\alpha)}(X^{(j)}).$$

However, in the case of Gibbs distributions with a single observation $X(\Lambda) = \{X_i : i \in \Lambda\}$, an empirical estimate of the covariance in (2.29) is not straightforward. But an approximate estimate of the covariance may be derived as follows: Let π_{θ_0} be the underlying true (infinite-volume) Gibbs distribution and suppose that

$$\lim_{\Lambda \to \mathscr{Z}^d} \frac{1}{\Lambda} \operatorname{Cov}_{\theta_0,\Lambda,z}\left(x(\Lambda) \cdot \nabla U_{\Lambda,z}^{(\beta)}, G_\Lambda^{(\alpha)}\right)$$

$$= \sum_{j\in\mathscr{Z}^d} \left\{ \left\langle \left(x_0 \frac{\partial \mathscr{U}_0^{(\beta)}}{\partial x_0} \right) G_j^{(\alpha)} \right\rangle^{(\theta_0)} - \left\langle x_0 \frac{\partial \mathscr{U}_0^{(\beta)}}{\partial x_0} \right\rangle^{(\theta_0)} \left\langle G_j^{(\alpha)} \right\rangle^{(\theta_0)} \right\}. \tag{2.31}$$

Typically such an equation holds, and its right-hand side is finite, if π_{θ_0} is

translation invariant and has good mixing properties. If it holds, we estimate

$$\left\langle x_0 \frac{\partial \mathscr{U}_0^{(\beta)}}{\partial x_0} \right\rangle^{(\theta_0)} \sim \bar{A}_\Lambda^{(\beta)} = \frac{1}{|\Lambda|} \sum_{i \in \Lambda} x_i \frac{\partial \mathscr{U}_i^{(\beta)}}{x_i},$$

$$\left\langle G_j^{(\alpha)} \right\rangle^{(\theta_0)} \sim \bar{G}_\Lambda^{(\alpha)} = \frac{1}{|\Lambda|} \sum_{i \in \Lambda} G_i^{(\alpha)}$$

and approximate the sum over $\mathscr{Z}^d$ by a sum over some neighborhood S_0 of $0 \in \mathscr{Z}^d$, for example, $S_0 = \{0\} \cup \mathscr{N}_0$. Then we arrive at the following approximation for the covariance:

$$\mathrm{Cov}_{\theta, \Lambda, z}\left(x(\Lambda) \cdot \nabla U_{\Lambda, z}^{(\beta)}, G_\Lambda^{(\alpha)}\right) \sim \sum_{i \in \Lambda} \left\{ \left[x_i \frac{\partial \mathscr{U}_i^{(\beta)}}{\partial x_i} - \bar{A}_\Lambda^{(\beta)} \right] \sum_{j \in S_i} \left[G_j^{(\alpha)} - \bar{G}_\Lambda^{(\alpha)} \right] \right\},$$

where $S_i = \tau^i S_0$, and τ^i denotes translation on $\mathscr{Z}^d$.

REMARK. For iid random variables, the second version of the VM is computationally as easy as the first. Both versions have the same order of accuracy. This is not true in general for Gibbs distributions, because the preceding approximations of the covariance in (2.29) may not be accurate if the underlying true distribution does not have good mixing properties.

C. *The variational method for incomplete data.* Here we extend the variational procedure of the previous subsection to the case of incomplete data. For simplicity, we treat only degraded data arising from additive noise. More specifically, we assume that at each pixel $i \in \Lambda$ we observe $y_i = x_i + \eta_i$, $x_i \in \mathbb{R}^n$, $\eta_i \in \mathbb{R}^n$, where $\{\eta_i\}$ are iid (independent of x_i) with probability law $Q(\eta_i) d\eta_i$. The marginal of $y = y(\Lambda) = \{y_i : i \in \Lambda\}$ is

$$P_{\theta, \Lambda}(y(\Lambda)) = \int_{\Omega_\Lambda} \pi_{\theta, \Lambda}(x(\Lambda)) Q(y(\Lambda) - x(\Lambda))\, dx(\Lambda)$$

$$= \int_{\Omega_\Lambda} \pi_{\theta, \Lambda}(y(\Lambda) - \eta(\Lambda)) Q(\eta(\Lambda))\, d\eta(\Lambda),$$

where $Q(\eta(\Lambda)) = \prod_{i \in \Lambda} Q(\eta_i)$ and $d\eta(\Lambda) = \prod_{i \in \Lambda} d\eta_i$. Here and below the presence of b.c. z is suppressed. In the remainder of this subsection the volume Λ will be fixed, and its presence will be dropped. Thus x, y, η will stand for $x(\Lambda), y(\Lambda), \eta(\Lambda)$, respectively. As in the complete data case we choose m vector statistics $\mathbf{W}^{(\alpha)}(y)$, $\alpha = 1, \ldots, m$, so that

$$\int \nabla \cdot \left(\mathbf{W}^{(\alpha)}(y) P_\theta(y) \right) dy = 0, \qquad \alpha = 1, \ldots, m.$$

This leads to the identity

$$\int \{ \mathbf{W}^{(\alpha)}(y) \cdot \nabla \ln P_\theta(y) \} P_\theta(dy) = -\int \nabla \cdot \mathbf{W}^{(\alpha)}(y) P_\theta(dy). \tag{2.32}$$

A straightforward computation gives

$$\nabla \ln P_\theta(y) = -\sum_{\beta=1}^{m} \theta^{(\beta)} E_\theta(\nabla U^{(\beta)}|y),$$

where $E_\theta(\cdot|y)$ denotes conditional expectation with respect to the posterior $P_\theta(dx|y)$. Thus

$$\sum_{\beta=1}^{m} \theta^{(\beta)} \int \{\mathbf{W}^{(\alpha)}(y) \cdot E_\theta(\nabla U^{(\beta)}|y)\} P_\theta(dy) = \int \{\nabla \cdot \mathbf{W}^{(\alpha)}(y)\} P_\theta(dy), \tag{2.33}$$

$$\alpha = 1, \ldots, m.$$

Now we replace the theoretical expectations [with respect to $P_\theta(dy)$] by the empirical expectations. If $Y = Y(\Lambda) = \{Y_i\colon i \in \Lambda\}$ is a single observation in Λ, we obtain

$$\sum_{\beta=1}^{m} \theta^{(\beta)} \{\mathbf{W}^{(\alpha)}(Y) \cdot E_\theta(\nabla U^{(\beta)}|Y)\} = \nabla \cdot \mathbf{W}^{(\alpha)}(Y), \qquad \alpha = 1, \ldots, m. \tag{2.34}$$

Good choices for $\mathbf{W}^{(\alpha)}(y)$ depend on the particular applications. The variational equations (2.34) are nonlinear, but they are simpler than the ML equations. They may be solved by an iterative procedure such as Newton's method or variants of it. An equivalent form of (2.34) is obtained by noting that $\nabla \ln P_\theta(y) = E_\theta(\nabla \ln Q(x - y)|y)$. Inserting this in (2.32) and replacing again theoretical expectations by empirical expectations we arrive at ($\alpha = 1, \ldots, m$)

$$\mathbf{W}^{(\alpha)}(Y) \cdot E_\theta(\nabla \ln Q(x - Y)|Y) = -\nabla \cdot \mathbf{W}^{(\alpha)}(Y). \tag{2.35}$$

In Section 4 we test (2.35) with an example of iid random variables governed by an exponential family.

The second version of the VM for incomplete data proceeds as in the case of complete data, and for this reason, we will not spell out its details here.

3. Consistency and asymptotic normality results. In this section we prove that the VE corresponding to (2.24) exists and is consistent and asymptotically normal as $\Lambda \to \mathscr{Z}^d$, under suitable conditions. Asymptotic normality holds if the underlying true distribution is ergodic; if it is not ergodic but only translation invariant, then we establish an asymptotic law which, in general, is not normal. Our asymptotic normality result involves a new central limit theorem (Theorem 3.3) which is of independent interest. Throughout this section the potentials $\Phi^{(\alpha)}$, $\alpha = 1, \ldots, m$, in addition to being superstable and regular, will be assumed to have continuous first and second derivatives, and to be of finite range (with interaction radius R_0).

3.1. *Preliminaries and main results.* Let $\Lambda \subset \mathscr{Z}^d$ be a finite volume in $\mathscr{Z}^d$. The (exterior) boundary $\partial\Lambda$ of Λ has been defined in Section 2B, and we

set $\overline{\Lambda} = \Lambda \cup \partial\Lambda$. We replace (2.24) by

$$(3.1)\quad \sum_{\beta=1}^{m} \theta^{(\beta)} \frac{1}{|\Lambda|} \sum_{i\in\Lambda} \frac{\partial \mathscr{U}_i^{(\alpha)}}{\partial x_i} \frac{\partial \mathscr{U}_i^{(\beta)}}{\partial x_i} = \frac{1}{|\Lambda|} \sum_{i\in\Lambda} \frac{\partial^2 \mathscr{U}_i^{(\alpha)}}{\partial x_i^2}, \qquad \alpha = 1, \ldots, m,$$

where $\mathscr{U}_i^{(\alpha)} = \mathscr{U}_i^{(\alpha)}(x_i, x(\mathscr{N}_i))$, $\alpha = 1, \ldots, m$, is defined as in (2.12a). Here, in contrast to (2.24), we assume that we are given the data $x(\overline{\Lambda}) = \{x_i\colon i \in \overline{\Lambda}\}$ in $\overline{\Lambda} = \Lambda \cup \partial\Lambda$ rather than in Λ (in this section the data are generic and will be denoted by lower case letters). Thus (3.1) contains only the observed data $x(\overline{\Lambda})$, and no boundary conditions enter in (3.1). We are interested in estimating the parameter vector θ from a single observation $x(\overline{\Lambda})$, and then studying consistency and asymptotic properties as $\Lambda \to \mathscr{Z}^d$. The sequence (or net) of observations $x(\overline{\Lambda})$ for an expanding sequence (net) of volumes $\Lambda \subset \mathscr{Z}^d$ is assumed to arise from an infinite sample $x \in \Omega$, $x = \{x_i\colon i \in \mathscr{Z}^d\}$, for which we observe larger and larger pieces $x(\overline{\Lambda}) = x|_{\overline{\Lambda}}$.

In order to prove consistency, we need an *identifiability condition*. The natural identifiability condition would be: $\theta_0 \in \Theta$ is identifiable if

$$(3.2)\qquad \theta \neq \theta_0, \qquad \theta \in \Theta \quad \text{implies} \quad G_0(\theta) \cap G_0(\theta_0) = \varnothing.$$

However, we will impose an alternative identifiability condition: $\theta_0 \in \Theta$ is identifiable if for $C = (C^{(1)}, \ldots, C^{(m)}) \in \mathbb{R}^m$,

$$(3.3)\qquad \sum_{\alpha=1}^{m} C^{(\alpha)} \mathscr{U}_0^{(\alpha)}(x_0, x(\mathscr{N}_0)) = \text{constant},$$

$$\pi_{\theta_0}\text{-a.s. for all } \pi_{\theta_0} \in G_0(\theta_0), \text{ iff } C = 0.$$

This condition is in general stronger than (3.2) [it implies (3.2), but (3.2) does not necessarily implies (3.3)]. Since $\mathscr{U}_0^{(\alpha)}(x_i, x(\mathscr{N}_i))$ and $\mathscr{U}_0^{(\alpha)}(x_i, x(\mathscr{N}_i)) - \mathscr{U}_0^{(\alpha)}(0, x(\mathscr{N}_i))$ give rise to the same local characteristics (2.15b) for all $\theta \in \Theta$ and $i \in \mathscr{Z}^d$, we can normalize the potentials so that $\mathscr{U}_0^{(\alpha)}(x_i, x(\mathscr{N}_i)) = 0$ whenever $x_i = 0$. Assuming this normalization, and the differentiability of the potentials, condition (3.3) is equivalent to

$$(3.3')\quad \sum_{\alpha=1}^{m} C^{(\alpha)} \frac{\partial \mathscr{U}_0^{(\alpha)}}{\partial x_0} = 0, \qquad \pi_{\theta_0}\text{-a.s. for all } \pi_{\theta_0} \in G_0(\theta_0), \text{ iff } C = 0.$$

We will use this form of the identifiability condition. We will also need a precise analogue of the divergence theorem (2.17), in the limit $\Lambda = \mathscr{Z}^d$. We will assume

$$(3.4)\qquad \int_\Omega d\pi_{\theta_0} \int_{\mathbb{R}^n} \left\{ \frac{\partial}{\partial x_0} \left[\frac{\partial \mathscr{U}_0^{(\alpha)}}{\partial x_0} \pi_{\theta_0}(x_0 | x(\mathscr{N}_0)) \right] \right\} dx_0 = 0$$

$$\text{for all } \pi_{\theta_0} \in G_0(\theta_0), \ \alpha = 1, \ldots, m.$$

Our existence and consistency result, to be proven in Section 3.2, is the following:

THEOREM 3.1 (Existence and consistency). *Let $\theta_0 \in \Theta$ be the true parameter vector, and $\pi_{\theta_0} \in G_0(\theta_0)$ any translation invariant Gibbs distribution. Assume that* (3.3′) *and* (3.4) *hold, and that $(\partial \mathscr{U}_0^{(\alpha)}/\partial x_0)^2$ and $\partial^2 \mathscr{U}_0^{(\alpha)}/\partial x_0^2$ are π_{θ_0}-integrable. Then for sufficiently large Λ,* (3.1) *has a unique solution $\hat{\theta}_\Lambda$ which π_{θ_0}-a.s. converges to θ_0 as $\Lambda \to \mathscr{Z}^d$.*

REMARK 3.1.1. The limit $\Lambda \to \mathscr{Z}^d$ in Theorem 3.1 and throughout this paper will be taken in the sense of van Hove [38].

REMARK 3.1.2. Because of the regularity (2.8) of the Gibbs distributions, the integrability assumptions in the theorem are easily satisfied in practice.

REMARK 3.1.3. The assumption $\pi_{\theta_0} \in G_0(\theta_0)$ [rather than $\pi_{\theta_0} \in G(\theta_0)$] could possibly be eliminated using large deviations results as in [7].

Our asymptotic normality result, to be proven in Section 3.3, is the following:

THEOREM 3.2 (Asymptotic normality). *Let θ_0 be the true parameter vector, and π_{θ_0} any ergodic Gibbs distribution. Assume* (3.3′) *and*

$$\int_{\mathbb{R}^n} \frac{\partial}{\partial x_0}\left[\frac{\partial \mathscr{U}_0^{(\alpha)}}{\partial x_0}\pi_{\theta_0}(x_0|x(\mathscr{N}_0))\right] dx_0 = 0, \qquad \pi_{\theta_0}\text{-a.s.} \quad \alpha = 1, \ldots, m. \tag{3.5}$$

Also assume that $\partial \mathscr{U}_0^{(\alpha)}/\partial x_0$ and $\partial^2 \mathscr{U}_0^{(\alpha)}/\partial x_0^2$ are in $L_k(d\pi_{\theta_0})$ for any positive integer k. Then the VE of Theorem 3.1 *is asymptotically normal, that is, under π_{θ_0}*

$$\sqrt{|\Lambda|}\,(\hat{\theta}_\Lambda - \theta_0) \to_{\mathscr{D}} N(0, \Sigma), \tag{3.6}$$

where Σ is an explicitly computable covariance matrix (see Section 3.3), *and $\to_{\mathscr{D}}$ denotes convergence in distribution.*

REMARK 3.2.1. If π_{θ_0} is not ergodic but only translation invariant, then we prove at the end of Section 3.3, that $\sqrt{|\Lambda|}\,(\hat{\theta}_\Lambda - \theta_0)$ converges in distribution to a nonnormal law.

REMARK 3.2.2. Condition (3.5) is a stronger form of (3.4); our proof of Theorem 3.2 uses a slightly weaker form of (3.5).

REMARK 3.2.3. Because of the regularity property (2.8) of Gibbs distributions, the integrability conditions of Theorem 3.2 are easily satisfied in practice.

3.2. *Proof of Theorem* 3.1. Throughout this subsection $\mathscr{S}$ will denote the σ-field generated by the translation invariant (measurable) subsets of $\Omega = (R^n)^{\mathscr{Z}^d}$, and $E_{\theta_0}(\cdot|\mathscr{S})$ will denote conditional expectation with respect to π_{θ_0}. The proof of the theorem will be given via four lemmas.

LEMMA 3.1. *Let*

$$\hat{T}_\Lambda^{(\alpha,\beta)}(x) = \hat{T}_\Lambda^{(\alpha,\beta)}\big(x(\bar{\Lambda})\big) = \frac{1}{|\Lambda|}\sum_{i\in\Lambda}\frac{\partial \mathscr{U}_i^{(\alpha)}}{\partial x_i}\frac{\partial \mathscr{U}_i^{(\beta)}}{\partial x_i}. \tag{3.7}$$

Under the hypothesis of Theorem 3.1 *we have* π_{θ_0}*-a.s.,*

$$\begin{aligned}
&\hat{T}_\Lambda^{(\alpha,\beta)}(x) \xrightarrow{\Lambda\to\mathscr{Z}^d} T^{(\alpha,\beta)}(\cdot) = E_{\theta_0}\left\{\frac{\partial \mathscr{U}_0^{(\alpha)}}{\partial x_0}\frac{\partial \mathscr{U}_0^{(\beta)}}{\partial x_0}\middle|\mathscr{S}\right\},\\
&\frac{1}{|\Lambda|}\sum_{i\in\Lambda}\frac{\partial^2 \mathscr{U}_i^{(\alpha)}}{\partial x_i^2} \xrightarrow{\Lambda\to\mathscr{Z}^d} E_{\theta_0}\left\{\frac{\partial^2 \mathscr{U}_0^{(\alpha)}}{\partial x_0^2}\middle|\mathscr{S}\right\}.
\end{aligned} \tag{3.8}$$

PROOF. This is a straightforward application of the ergodic theorem, noting that $|\bar{\Lambda}|/|\Lambda| \to 1$ as $\Lambda \to \mathscr{Z}^d$. □

REMARK. If π_{θ_0} is ergodic, then $T^{(\alpha,\beta)}(\cdot)$ is a constant.

LEMMA 3.2. *Under the hypothesis of Theorem* 3.1, *the matrix* $T^{(\alpha,\beta)}(\cdot)$ *is* π_{θ_0}*-a.s. positive-definite, and satisfies*

$$\sum_{\beta=1}^{m} T^{(\alpha,\beta)}(\cdot)\theta_0^{(\beta)} = E_{\theta_0}\left\{\frac{\partial^2 \mathscr{U}_0^{(\alpha)}}{\partial x_0^2}\middle|\mathscr{S}\right\}.$$

PROOF. First, we prove the lemma for π_{θ_0} ergodic. Then an application of the ergodic decomposition (2.9) will yield the lemma for any $\pi_{\theta_0} \in G_0(\theta_0)$.

Note that for any $C = \{C^{(\alpha)}\}_{\alpha=1}^m \in \mathbb{R}^m$,

$$\sum_{\alpha=1}^{m} T^{(\alpha,\beta)}C^{(\alpha)}C^{(\beta)} = E_{\theta_0}\left\{\left(\sum_{\alpha=1}^{m} C^{(\alpha)}\frac{\partial \mathscr{U}_0^{(\alpha)}}{\partial x_0}\right)^2\right\} \geq 0$$

with equality if and only if

$$C^{(\alpha)}\frac{\partial \mathscr{U}_0^{(\alpha)}}{\partial x_0} = 0, \qquad \pi_{\theta_0}\text{-a.s.},$$

which, by (3.3′), implies $C = 0$. Next, using the Markov property and (3.4), we

obtain

$$E_{\theta_0}\left\{\frac{\partial^2 \mathscr{U}_0^{(\alpha)}}{\partial x_0^2}\right\} = \int_\Omega \pi_{\theta_0}(dx)\int_{\mathbb{R}^n}\frac{\partial}{\partial x_0}\left[\frac{\partial \mathscr{U}_0^{(\alpha)}}{\partial x_0}\pi_{\theta_0}(x_0|x(\mathscr{N}_0))\right]dx_0$$

$$+\int_\Omega \pi_{\theta_0}(dx)\left[\frac{\partial \mathscr{U}_0^{(\alpha)}}{\partial x_0}\frac{\partial H_0^{(\theta_0)}}{\partial x_0}\right]$$

$$= \sum_{\beta=1}^m \theta^{(\beta)} E_{\theta_0}\left\{\frac{\partial \mathscr{U}_0^{(\alpha)}}{\partial x_0}\frac{\partial \mathscr{U}_0^{(\beta)}}{\partial x_0}\right\}.$$

This completes the proof of the lemma when π_{θ_0} is ergodic.

Next assume that $\pi_{\theta_0} \in G_0(\theta_0)$ is not ergodic. Let $\Psi(x) = \{\Psi^{(\alpha)}(x)\}_{\alpha=1}^m$ be $\mathscr{S}$-measurable. To show that $T^{(\alpha,\beta)}(\cdot)$ is positive-definite, we will show that

$$\sum_{\alpha,\beta=1}^m T^{(\alpha,\beta)}(\cdot)\Psi^{(\alpha)}(x)\Psi^{(\beta)}(x) = 0 \qquad \pi_{\theta_0}\text{-a.s.},$$

implies $\Psi(x) = 0$, π_{θ_0}-a.s. Integrating over Ω and using the ergodic decomposition (2.9), we obtain

$$\int_{\mathscr{E}_0(\Phi)}\rho_{\pi_{\theta_0}}(d\xi)\int_\Omega\left[\sum_{\alpha=1}^m \Psi^{(\alpha)}(x)\frac{\partial \mathscr{U}_0^{(\alpha)}}{\partial x_0}\right]^2 P_{\theta_0}^{(\xi)}(dx) = 0.$$

Since $P_{\theta_0}^{(\xi)}$ is ergodic and $\Psi^{(\alpha)}(x)$ is $\mathscr{S}$-measurable, we have that $\Psi^{(\alpha)}(x) = C^{(\alpha)}$, $P_{\theta_0}^{(\xi)}$-a.s. Thus the equation above implies

$$\sum_{\alpha=1}^m C^{(\alpha)}\frac{\partial \mathscr{U}_0^{(\alpha)}}{\partial x_0} = 0 \qquad P_{\theta_0}^{(\xi)}\text{-a.s.},$$

and by (3.3′), $C = \{C^{(\alpha)}\}_{\alpha=1}^m = 0$. Thus

$$\pi_{\theta_0}\{x\colon \Psi(x) \neq 0\} = \int \rho_{\pi_{\theta_0}}(d\xi)P_{\theta_0}^{(\xi)}\{x\colon \Psi(x) \neq 0\} = 0.$$

Next we show that $T^{(\alpha,\beta)}(\cdot)$ satisfies the equation of the lemma: Let $A \in \mathscr{S}$, then

$$\int_A E_{\pi_{\theta_0}}\left\{\frac{\partial^2 \mathscr{U}_0^{(\alpha)}}{\partial x_0^2}\middle|\mathscr{S}\right\}\pi_{\theta_0}(dx) = \int_A \frac{\partial^2 \mathscr{U}_0^{(\alpha)}}{\partial x_0^2}\pi_{\theta_0}(dx).$$

Using (2.9), Fubini's theorem, and (3.4) we have

$$\int_A E_{\pi_{\theta_0}}\left\{\frac{\partial^2 \mathscr{U}_0^{(\alpha)}}{\partial x_0^2}\middle|\mathscr{S}\right\}\pi_{\theta_0}(dx) = \int_{\mathscr{E}_0(\Phi)}\rho_{\pi_{\theta_0}}(d\xi)\int_A E_{P_{\theta_0}^{(\xi)}}\left\{\frac{\partial^2 \mathscr{U}_0^{(\alpha)}}{\partial x_0^2}\right\}P_{\theta_0}^{(\xi)}(dx)$$

$$= \int_{\mathscr{E}_0(\Phi)}\rho_{\pi_{\theta_0}}(d\xi)\int_A E_{P_{\theta_0}^{(\xi)}}\left\{\frac{\partial \mathscr{U}_0^{(\alpha)}}{x_0}\frac{\partial H_0^{(\theta_0)}}{\partial x_0}\right\}P_{\theta_0}^{(\xi)}(dx)$$

$$= \int_A E_{\pi_{\theta_0}}\left\{\frac{\partial \mathscr{U}_0^{(\alpha)}}{x_0}\frac{\partial H_0^{(\theta_0)}}{\partial x_0}\right\}\pi_{\theta_0}(dx).$$

This proves the lemma, since $A \in \mathscr{S}$ is arbitrary. □

LEMMA 3.3. *Let* $A = (a^{(\alpha,\beta)})$ *be an* $m \times m$ *nonsingular matrix, and* $A_n = (a_n^{(\alpha,\beta)})$ *be a sequence of* $m \times m$ *matrices, which converges to* A *entry-wise, that is,* $a_n^{(\alpha,\beta)} \to a^{(\alpha,\beta)}$. *Then there exists an* N_0 *such that for* $n \geq N_0$, A_n *is invertible, and* $A_n^{-1} \to A^{-1}$ *in norm as* $n \to \infty$.

PROOF. Let $\|\cdot\|$ denote any of equivalent norms on the space of $m \times m$ matrices. We have $\|A_n - A\| \to 0$. Let N_0 be such that $\|A_n - A\| < (\|A^{-1}\|)^{-1}$ for all $n \geq N_0$. Thus $\|A^{-1}(A - A_n)\| < \|A^{-1}\| \, \|A - A_n\| < 1$. Therefore $I - A^{-1}(A - A_n)$ and $A_n = A[I - A^{-1}(A - A_n)]$ are invertible for $n \geq N_0$. Using the series expansion of $[I - A^{-1}(A - A_n)]^{-1}$ we obtain for $n \geq N_0$,

$$\|A_n^{-1} - A^{-1}\| \leq \|A^{-1}\| \frac{\|A^{-1}\| \, \|A - A_n\|}{1 - \|A^{-1}\| \, \|A - A_n\|}.$$

Hence $\|A_n^{-1} - A^{-1}\| \to 0$ as $n \to \infty$. □

LEMMA 3.4. *Let* $A = (a^{(\alpha,\beta)})$ *be an* $m \times m$ *nonsingular matrix and consider the system of equations*

$$\sum_{\beta=1}^{m} a^{(\alpha,\beta)}\theta^{(\beta)} = b^{(\alpha)}, \qquad \alpha = 1, \ldots, m. \tag{3.9a}$$

Let

$$\sum_{\beta=1}^{m} a_n^{(\alpha,\beta)}\theta_n^{(\beta)} = b_n^{(\alpha)}, \qquad \alpha = 1, \ldots, m \tag{3.9b}$$

be a sequence of linear systems so that $a_n^{(\alpha,\beta)} \to a^{(\alpha,\beta)}$, *and* $b_n^{(\alpha)} \to b^{(\alpha)}$, $\alpha = 1, \ldots, m$, $\beta = 1, \ldots, m$. *Then there exists an* N_0 *such that for* $n \geq N_0$ (3.9b) *has a unique solution* $\theta_n = (\theta_n^{(1)}, \ldots, \theta_n^{(m)})$ *such that* $|\theta_n - \theta| \to 0$ *as* $n \to \infty$.

PROOF. By Lemma 3.3 there exists an N_0 such that for $n \geq N_0$, A_n is invertible. Thus for $n \geq N_0$, (3.9b) has a unique solution

$$\theta_n = A_n^{-1}b_n = A_n^{-1}(b_n - b) + \left(A_n^{-1} - A^{-1}\right)b + \theta,$$

where θ is the unique solution of (3.9a). This, together with the fact that $\|A_n^{-1} - A^{-1}\| \to 0$, quickly yields the lemma. □

PROOF OF THEOREM 3.1. The proof is now a straightforward consequence of Lemmas 3.1–3.4. □

3.3. *Proof of Theorem* 3.2. Subtracting

$$\sum_{\beta=1}^{m} \theta_0^{(\beta)} \frac{1}{|\Lambda|} \sum_{i \in \Lambda} \frac{\partial \mathscr{U}_i^{(\alpha)}}{\partial x_i} \frac{\partial \mathscr{U}_i^{(\beta)}}{\partial x_i}$$

from both sides of (3.1) and multiplying by $\sqrt{|\Lambda|}$ we obtain,

$$\sum_{\beta=1}^{m} \hat{T}_{\Lambda}^{(\alpha,\beta)}(x)\sqrt{|\Lambda|}\left(\hat{\theta}_{\Lambda}^{(\beta)} - \theta_0^{(\beta)}\right)$$

$$= \frac{1}{\sqrt{|\Lambda|}} \sum_{i\in\Lambda} \left\{ \frac{\partial^2 \mathscr{U}_i^{(\alpha)}}{\partial x_i^2} - \frac{\partial \mathscr{U}_i^{(\alpha)}}{\partial x_i} \frac{\partial H_i^{(\theta_0)}}{\partial x_i} \right\}, \qquad \alpha = 1, \ldots, m, \tag{3.10}$$

where $\hat{T}_{\Lambda}^{(\alpha,\beta)}(x)$ has been defined in (3.5). The limit of this matrix is given by (3.6). Below we will show that the right-hand side of (3.10) satisfies a central limit theorem. Let

$$Y_i^{(\alpha)} = \frac{\partial^2 \mathscr{U}_i^{(\alpha)}}{\partial x_i^2} - \frac{\partial \mathscr{U}_i^{(\alpha)}}{\partial x_i} \frac{\partial H_i^{(\theta_0)}}{\partial x_i},$$

$$Y_{\Lambda}^{(\alpha)} = \frac{1}{\sqrt{|\Lambda|}} \sum_{i\in\Lambda} Y_i^{(\alpha)},$$

$$\mathbf{Y}_{\Lambda} = \left\{Y_{\Lambda}^{(\alpha)}\right\}_{\alpha=1}^{m}.$$

We will prove the following results:

THEOREM 3.3. *Under the conditions of Theorem* 3.2, *for any* $\mathbf{t} \in \mathbb{R}^m$ *we have*

$$\lim_{\Lambda \to \mathscr{Z}^d} E_{\theta_0}\{\exp[\mathbf{it} \cdot \mathbf{Y}_{\Lambda}]\} = \exp\left[-\tfrac{1}{2}\mathbf{t} \cdot B\mathbf{t}\right], \tag{3.11}$$

where E_{θ_0} *denotes expectation with respect to* π_{θ_0}, *and the matrix* $B = (B^{(\alpha,\beta)})$ *is given by*

$$B^{(\alpha,\beta)} = E_{\theta_0}\left\{ Y_0^{(\alpha)} \sum_{j\in V_0} Y_j^{(\beta)} \right\}, \tag{3.12}$$

$$V_0 = \{0\} \cup \mathscr{N}_0.$$

In particular, under π_{θ_0},

$$\mathbf{Y}_{\Lambda} \to_{\mathscr{D}} N(0, B). \tag{3.13}$$

REMARK 3.3.1. The proof of Theorem 3.3 is lengthy and technical, but property (3.15) is the most intrinsic property needed for the validity of the theorem. In fact, any set of dependent random variables $\{Y_j^{(\alpha)}\}$, $j \in \Lambda$, $\alpha = 1, \ldots, m$, that have property (3.15) and satisfy appropriate integrability conditions, satisfy (3.11).

REMARK 3.3.2. The matrix Σ in (3.5) is given by $\Sigma = T^{-1}BT$ where T is the (constant) matrix in (3.7).

REMARK 3.3.3. We believe that (3.11) still holds if the interactions are not of finite range, but satisfy

$$\sum_{j \in \mathscr{Z}^d} E_{\theta_0}\{Y_0^{(\alpha)} Y_j^{(\beta)}\} < \infty.$$

REMARK 3.3.4. Theorem 3.3, together with the π_{θ_0}-a.s. convergence of the matrix $\hat{T}^{(\alpha,\beta)}(x)$, and Lemma 6.4.1 of [28], page 439, imply Theorem 3.2. Thus we need only to prove Theorem 3.3; its proof is based on the following proposition.

PROPOSITION 3.1. *Under the hypotheses of Theorem* 3.2, *we have for any integer* $k \geq 0$,

$$\lim_{\Lambda \to \mathscr{Z}^d} E_{\theta_0}\{(\mathbf{t} \cdot \mathbf{Y}_\Lambda)^k\} = \int (\mathbf{t} \cdot \mathbf{y})^k F(d\mathbf{y}), \tag{3.14}$$

where F *is the distribution with characteristic function* $\exp\{-(1/2)\mathbf{t} \cdot B\mathbf{t}\}$.

The proof of this proposition is technical and lengthy. It will use a series of lemmas.

LEMMA 3.5. *For any* $\pi_{\theta_0} \in G(\theta_0)$, *we have*

$$E_{\theta_0}\{Y_j^{(\alpha)} | x(\mathscr{Z}^d - \{j\})\} = 0, \qquad j \in \mathscr{Z}^d,\ \alpha = 1, \ldots, m. \tag{3.15}$$

PROOF. Consider

$$E_{\theta_0}\left\{ \frac{\partial^2 \mathscr{U}_j^{(\alpha)}}{\partial x_j^2} \middle| x(\mathscr{Z}^d - \{j\}) \right\} = \int_{\mathbb{R}^n} \frac{\partial^2 \mathscr{U}_j^{(\alpha)}}{\partial x_j^2} \pi_{\theta_0}(x_j | x(\mathscr{Z}^d - \{j\}))\, dx_j$$

and apply condition (3.4). □

Let $Z_j = \mathbf{t} \cdot \mathbf{Y}_j = \sum_{\alpha=1}^m t^{(\alpha)} Y_j^{(\alpha)}$, $j \in \Lambda$, and write

$$E_{\theta_0}\{(\mathbf{t} \cdot \mathbf{Y}_\Lambda)^k\} = \frac{1}{|\Lambda|^{k/2}} \sum_{j_1, \ldots, j_k \in \Lambda} E_{\theta_0}\{Z_{j_1} \cdots Z_{j_k}\}. \tag{3.16}$$

We will split the sum in (3.16) into three parts according to the "clustering" (separation) of the pixels in the k-tuple $(j_1, \ldots, j_k) \in \Lambda^k$ ($\Lambda^k = \Lambda \times \cdots \times \Lambda$, k copies of Λ). The splitting is facilitated by the following definitions.

DEFINITIONS. Recall that R_0 denotes the interaction radius of the potentials.

1. The distance between two sets of pixels $\{i_1, \ldots, i_p\}$ and $\{j_1, \ldots, j_q\}$ is defined by

$$D(\{i_1, \ldots, i_p\}, \{j_1, \ldots, j_q\}) = \min\{|i_l - j_m| : l = 1, \ldots, p,\ m = 1, \ldots, q\}.$$

2. A k-tuple $(j_1, \ldots, j_k)$ is said to be an R_0-cluster if it cannot be separated into two subsets with distance strictly larger than R_0.
3. The number of R_0-clusters of a k-tuple $(j_1, \ldots, j_k)$ will be denoted by

$$K_{R_0} = K_{R_0}(j_1, \ldots, j_k).$$

Next we define three disjoint subsets B_1, B_2, B_3 such that $\Lambda^k = B_1 \cup B_2 \cup B_3$: (a) B_1 is the set of k-tuples $(j_1, \ldots, j_k) \in \Lambda^k$ which have at least one R_0-cluster containing only a single pixel. (Note that the complement, $\Lambda^k \setminus B_1$, of B_1, contains at most $[k/2]$ R_0-clusters, where $[k/2]$ denotes the integer part of $k/2$.)

(b) B_2 is the subset of $\Lambda^k \setminus B_1$, defined by

$$B_2 = \left\{(j_1, \ldots, j_k) \in \Lambda^k \setminus B_1 \colon K_{R_0}(j_1, \ldots, j_k) < \frac{k}{2}\right\}.$$

(Note that $\Lambda^k = B_1 \cup B_2$ if k is odd; if k is even then there is one more subset.)

(c) B_3 is the complement of $B_1 \cup B_2$ in Λ^k, that is,

$$\Lambda^k = B_1 \cup B_2 \cup B_3, \qquad B_3 \cap (B_1 \cup B_2) = \varnothing$$

or equivalently,

$$B_3 = \left\{(j_1, \ldots, j_k) \in \Lambda^k \setminus (B_1 \cup B_2) \colon K_{R_0}((j_1, \ldots, j_k) = \frac{k}{2}\right\}.$$

[Note that B_3 is empty if k is odd; if k is even, say $k = 2l$, then any $(j_1, \ldots, j_k) \in B_3$ has exactly l R_0-clusters: Each cluster consists of two (possibly identical) pixels; the clusters are separated by a distance strictly larger than R_0.]

Now we write

$$E\{(\mathbf{t} \cdot \mathbf{Y}_\Lambda)^k\} = I^{(1)}_{k,\Lambda} + I^{(2)}_{k,\Lambda} + I^{(3)}_{k,\Lambda}, \tag{3.17a}$$

$$I^{(1)}_{k,\Lambda} = I^{(1)}_{k,\Lambda}(\mathbf{t}) = \frac{1}{|\Lambda|^{k/2}} \sum_{(j_1, \ldots, j_k) \subset B_1} E_{\theta_0}\{Z_{j_1} \cdots Z_{j_k}\}, \tag{3.17b}$$

$$I^{(2)}_{k,\Lambda} = I^{(2)}_{k,\Lambda}(\mathbf{t}) = \frac{1}{|\Lambda|^{k/2}} \sum_{(j_1, \ldots, j_k) \subset B_2} E_{\theta_0}\{Z_{j_1} \cdots Z_{j_k}\}, \tag{3.17c}$$

$$I^{(3)}_{k,\Lambda} = I^{(3)}_{k,\Lambda}(\mathbf{t}) = \frac{1}{|\Lambda|^{k/2}} \sum_{(j_1, \ldots, j_k) \subset B_3} E_{\theta_0}\{Z_{j_1} \cdots Z_{j_k}\}. \tag{3.17d}$$

Each of these terms is controlled (as $\Lambda \to \mathscr{Z}^d$) in the following three lemmas; the third lemma (Lemma 3.8) is the most subtle.

LEMMA 3.6.

$$\sum_{(j_1, \ldots, j_k) \in B_1} E_{\theta_0}\{Z_{j_1} \cdots Z_{j_k}\} = 0 \tag{3.18}$$

PROOF. Let j_1 form an R_0-cluster by itself. Using the Markov property and Lemma 3.5 we obtain

$$E_{\theta_0}\{Z_{j_1} \cdots Z_{j_k}\} = \int_\Omega E_{\theta_0}\{Z_{j_1} | x(\mathscr{Z}^d - \{j_1\})\} Z_{j_2} \cdots Z_{j_k} \pi_{\theta_0}(dx) = 0. \qquad \square$$

LEMMA 3.7.

$$\lim_{\Lambda \to \mathscr{Z}^d} \frac{1}{|\Lambda|^{k/2}} \sum_{(j_1, \ldots, j_k) \in B_2} E_{\theta_0}\{Z_{j_1} \cdots Z_{j_k}\} = 0. \tag{3.19}$$

PROOF. Since the number of R_0-clusters in B_2 is strictly less than $k/2$, it is easily seen that the number $|B_2|$, of k-tuples of B_2, satisfies $|B_2| \leq C|\Lambda|^{(k-1)/2}$ with a constant C independent of Λ (C depends on R_0 and k). Therefore as $\Lambda \to \mathscr{Z}^d$,

$$|I_{k\Lambda}^{(2)}| \leq \frac{1}{|\Lambda|^{k/2}} C|\Lambda|^{(k-1)/2} E_{\theta_0}\{|Z_0|^k\} \leq C|\mathbf{t}|^k \frac{1}{\sqrt{|\Lambda|}} E_{\theta_0}\{|\mathbf{Y}_0|^k\} \to 0. \qquad \square$$

REMARK. Lemmas 3.6 and 3.7 prove Proposition 3.1 when k is odd.

LEMMA 3.8. *Let k be a positive even integer. Then*

$$\lim_{\Lambda \to \mathscr{Z}^d} I_{k,\Lambda}^{(3)}(\mathbf{t}) = \frac{k!}{(k/2)!2^{k/2}} (\mathbf{t} \cdot B\mathbf{t})^{k/2},$$

where $I_{k,\Lambda}^{(3)}(\mathbf{t})$ is defined in (3.17d).

PROOF. By definition, any k-tuple $(j_1, \ldots, j_k) \in B_3$ has exactly $k/2$ R_0-clusters, and each cluster contains exactly two (possibly identical) pixels. This means that there exists a permutation π of $(1, 2, \ldots, k)$ so that the R_0-clusters of a k-tuple $(j_1, \ldots, j_k) \in B_3$ take the form $\{j_{\pi(2l-1)}, j_{\pi(2l)}\}$, $l = 1, \ldots, k/2$. By definition $|j_{\pi(2l-1)} - j_{\pi(2l)}| \leq R_0$, and the distance between any two clusters

$$\{j_{\pi(2l-1)}, j_{\pi(2l)}\} \quad \text{and} \quad \{j_{\pi(2m-1)}, j_{\pi(2m)}\}, \qquad m \neq l,$$

is strictly larger than R_0.

Let B_3^* be the subset of B_3 defined by

$$\begin{aligned} B_3^* = \{&((j_1, \ldots, j_k) \subset B_3 : |j_{2l-1} - j_{2l}| \leq R_0,\ l = 1, \ldots, k/2, \\ &\operatorname{dist}(\{j_{2l-1}, j_{2l}\}, \{j_{2m-1}, j_{2m}\}) \\ &> R_0,\ l \neq m,\ l, m = 1, 2, \ldots, k/2\}. \end{aligned} \tag{3.20}$$

Also, let $\mathscr{P}_k = \mathscr{P}(1, 2, \ldots, k)$ be the set of permutations of $(1, 2, \ldots, k)$. Then

it is easily seen that

$$\sum_{((j_1,\ldots,j_k)\in B_3} E_{\theta_0}\{Z_{j_1},\ldots,Z_{j_k}\}$$

$$= \frac{1}{(k/2)!2^{k/2}} \sum_{\pi\in\mathscr{P}_k} \sum_{(j_{\pi(1)},\ldots,j_{\pi(k)})\in B_3^*} E_{\theta_0}\left\{\prod_{l=1}^{k/2} Z_{j_{\pi(2l-1)}} Z_{j_{\pi(2l)}}\right\}. \tag{3.21}$$

Next we split the sum over B_3^* into two parts: Let $R > R_0$ and define two disjoint subsets $B_{31}^*(R)$ and $B_{32}^*(R)$ of B_3^* by

$$B_{31}^*(R) = \{(j_1,\ldots,j_k) \subset B_3^*: \text{at least two } R_0\text{-clusters have distance strictly less than } R\},$$

$$B_3^* = B_{31}^*(R) \cup B_{32}^*(R), \qquad B_{31}^*(R) \cap B_{32}^*(R) = \varnothing.$$

Note that the distance between any two R_0-clusters of a k-tuple $(j_1,\ldots,j_k) \in B_{32}^*(R)$ is larger than or equal to R. We write

$$\frac{1}{|\Lambda|^{k/2}} \sum_{(j_1,\ldots,j_k)\in B_3^*} E_{\theta_0}\left\{\prod_{l=1}^{k/2} Z_{j_{2l-1}} Z_{j_{2l}}\right\} = J_{k,\Lambda}^{(1)}(R) + J_{k,\Lambda}^{(2)}(R), \tag{3.22a}$$

where

$$J_{k,\Lambda}^{(1)}(R) = J_{k,\Lambda}^{(1)}(\mathbf{t}; R) = \frac{1}{|\Lambda|^{k/2}} \sum_{(j_1,\ldots,j_k)\in B_{31}^*(R)} E_{\theta_0}\left\{\prod_{l=1}^{k/2} Z_{j_{2l-1}} Z_{j_{2l}}\right\}, \tag{3.22b}$$

$$J_{k,\Lambda}^{(2)}(R) = J_{k,\Lambda}^{(2)}(\mathbf{t}; R) = \frac{1}{|\Lambda|^{k/2}} \sum_{(j_1,\ldots,j_k)\in B_{32}^*(R)} E_{\theta_0}\left\{\prod_{l=1}^{k/2} Z_{j_{2l-1}} Z_{j_{2l}}\right\}. \tag{3.22c}$$

The terms $J_{k,\Lambda}^{(1)}(R)$ and $J_{k,\Lambda}^{(2)}(R)$ are controlled in the next two lemmas.

LEMMA 3.9. *For any finite $R > R_0$, we have*

$$\lim_{\Lambda\to\mathscr{Z}^d} J_{k,\Lambda}^{(1)}(R) = 0. \tag{3.23}$$

PROOF. The proof is the same as in Lemma 3.7. □

LEMMA 3.10. *Given $\varepsilon > 0$ there exists an $R(\varepsilon) > R_0$ such that for $R \geq R(\varepsilon)$,*

$$\lim_{\Lambda\to\mathscr{Z}^d} |J_{k,\Lambda}^{(2)}(R) - (\mathbf{t}\cdot B\mathbf{t})^{k/2}| < \varepsilon \tag{3.24}$$

for $\mathbf{t}$ in any subset of $\mathbb{R}^m$.

PROOF. Note that

$$\mathbf{t}\cdot B\mathbf{t} = E_{\theta_0}\left\{Z_0 \sum_{j\in V_0} Z_j\right\}.$$

Recall that if $i \in \mathscr{Z}^d$, then $\mathscr{N}_i$ denotes the neighbors of i, and $V_i = \{i\} \cup \mathscr{N}_i$.

For any $R > R_0$ we define

$$B^{**}(R) = \left\{(j_1, \ldots, j_{k/2}) \in \Lambda^{k/2} : \operatorname{dist}(V_{i_l}, V_{i_m}) \geq R,\ l \neq m;\ l, m = 1, \ldots, \frac{k}{2}\right\}.$$

Then

$$J^{(2)}_{k,\Lambda}(R) = M^{(1)}_{k,\Lambda}(R) + M^{(2)}_{k,\Lambda}(R),$$

where

$$M^{(1)}_{k,\Lambda}(R) = \frac{1}{|\Lambda|^{k/2}} \sum_{((j_1, \ldots, j_{k/2})) \in B^{**}(R)} \left\{ E_{\theta_0}\left[\prod_{l=1}^{k/2} Z_{i_l} \sum_{i \in V_{i_l}} Z_i\right] - \prod_{l=1}^{k/2} E_{\theta_0}\left[Z_{i_l} \sum_{i \in V_{i_l}} Z_i\right]\right\},$$

$$M^{(2)}_{k,\Lambda}(R) = \frac{1}{|\Lambda|^{k/2}} \sum_{((j_1, \ldots, j_{k/2})) \in B^{**}(R)} \prod_{l=1}^{k/2} E_{\theta_0}\left[Z_{i_l} \sum_{i \in V_{i_l}} Z_i\right] - (\mathbf{t} \cdot B\mathbf{t})^{k/2}.$$

The ergodicity of π_{θ_0}, and the $L_k(d\pi_{\theta_0})$ integrability conditions of Theorem 3.2, imply that given $\varepsilon > 0$ there exists an $R(\varepsilon)$ sufficiently large so that for $R > R(\varepsilon)$,

$$\left| E_{\theta_0}\left[\prod_{l=1}^{k/2} Z_{i_l} \sum_{i \in V_{i_l}} Z_i\right] - \prod_{l=1}^{k/2} E_{\theta_0}\left[Z_{i_l} \sum_{i \in V_{i_l}} Z_i\right] \right| < \varepsilon$$

for $\mathbf{t}$ in compact sets of $\mathbb{R}^m$. Thus for sufficiently large R,

$$|M^{(1)}_{k,\Lambda}(R)| < \frac{|B^{**}(R)|}{|\Lambda|^{k/2}} \varepsilon.$$

By translation invariance,

$$\prod_{l=1}^{k/2} E_{\theta_0}\left[Z_{i_l} \sum_{i \in V_{i_l}} Z_i\right] = (\mathbf{t} \cdot B\mathbf{t})^{k/2}.$$

Therefore

$$|M^{(2)}_{k,\Lambda}(R)| = \left(1 - \frac{|B^{**}(R)|}{|\Lambda|^{k/2}}\right)(\mathbf{t} \cdot B\mathbf{t})^{k/2}.$$

Hence, for sufficiently large R,

$$|J^{(2)}_{k,\Lambda} - (\mathbf{t} \cdot B\mathbf{t})^{k/2}| \leq \frac{|B^{**}(R)|}{|\Lambda|^{k/2}} \varepsilon + \left(1 - \frac{|B^{**}(R)|}{|\Lambda|^{k/2}}\right)(\mathbf{t} \cdot B\mathbf{t})^{k/2}. \tag{3.25}$$

It is easily seen that

$$\frac{|B^{**}(R)|}{|\Lambda|^{k/2}} \to 1, \quad \text{as } \Lambda \to \mathscr{Z}^d.$$

This, together with (3.25), yields (3.24). □

PROOF OF LEMMA 3.8 (completed). Combining Lemmas 3.9 and 3.10 with the identity (3.21), and noting that the number of permutations in $\mathscr{P}_k = \mathscr{P}(1, \ldots, k)$ is $k!$, we quickly obtain Lemma 3.8. □

Lemmas 3.6–3.8 yield

$$\lim_{\Lambda \to \mathscr{Z}^d} E_{\theta_0}\{(\mathbf{t} \cdot \mathbf{Y}_\Lambda)^k\} = \begin{cases} 0, & \text{if } k \text{ is odd}, \\ \dfrac{k!}{(k/2)!2^{k/2}}(\mathbf{t} \cdot B\mathbf{t})^{k/2}, & \text{if } k \text{ is even}, \end{cases} \tag{3.26a}$$

$$= \int (\mathbf{t} \cdot y) F(dy). \tag{3.26b}$$

This proves Proposition 3.1. □

PROOF OF THEOREM 3.3. Note that the Taylor series of

$$\varphi_\Lambda(\mathbf{t}) = E_{\theta_0}\{\exp[\mathbf{i}\mathbf{t} \cdot \mathbf{Y}_\Lambda]\}$$

converges, as $\Lambda \to \mathscr{Z}^d$, term by term to the Taylor series of

$$\varphi(\mathbf{t}) = \exp\left[-\tfrac{1}{2}\mathbf{t} \cdot B\mathbf{t}\right].$$

Now we will prove that $\varphi_\Lambda(\mathbf{t})$ converges to $\varphi(\mathbf{t})$ as $\Lambda \to \mathscr{Z}^d$. Consider the Taylor series with a remainder

$$\varphi_\Lambda(\mathbf{t}) = E_{\theta_0}\{\cos(\mathbf{t} \cdot \mathbf{Y}_\Lambda) + \mathbf{i} \sin(\mathbf{t} \cdot \mathbf{Y}_\Lambda)\} = S_\Lambda^{(2N-1)} + R_\Lambda^{(2N)},$$

where

$$S_\Lambda^{(2N-1)} = S_\Lambda^{(2N-1)}(\mathbf{t}) = \sum_{k=0}^{2n-1} \frac{\mathbf{i}^k}{k!} E_{\theta_0}\{(\mathbf{t} \cdot \mathbf{Y}_\Lambda)^k\},$$

$$R_\Lambda^{(2N)} = R_\Lambda^{(2N)}(\mathbf{t})$$

$$= \frac{\mathbf{i}^{2N}}{(2N)!} E_{\theta_0}\{[\cos(\theta_1(\mathbf{t} \cdot \mathbf{Y}_\Lambda)) + \mathbf{i} \sin(\theta_2(\mathbf{t} \cdot \mathbf{Y}_\Lambda))](\mathbf{t} \cdot \mathbf{Y}_\Lambda)^{2N}\}$$

with $|\theta_1| \le 1$, $|\theta_2| \le 1$ (θ_1 and θ_2 are random quantities). Similarly $\varphi(\mathbf{t}) = S^{(2N-1)} + R^{(2N)}$. Note that

$$\begin{aligned}\limsup_{\Lambda \to \mathscr{Z}^d} |R_\Lambda^{(2N)}| &\le \sqrt{2}\,\frac{1}{(2N)!} \lim_{\Lambda \to \mathscr{Z}^d} E_{\theta_0}\left\{(\mathbf{t}\cdot \mathbf{Y}_\Lambda)^{2N}\right\} \\ &= \sqrt{2}\,\frac{1}{N!2^N}(\mathbf{t}\cdot B\mathbf{t})^N \\ &\to 0, \quad \text{as } N \to \infty.\end{aligned} \tag{3.27}$$

Also $S_\Lambda^{(2N-1)}$ converges (term by term) as $\Lambda \to \mathscr{Z}^d$ to $S^{(2N-1)}$. This, together with (3.27), implies the convergence of $\varphi_\Lambda(\mathbf{t})$ to $\varphi(\mathbf{t})$ as $\Lambda \to \mathscr{Z}^d$.

We end this section by deriving an asymptotic law for $\sqrt{|\Lambda|}\,(\hat{\theta}_\Lambda - \theta_0)$ when the true distribution π_{θ_0} is not ergodic but only translation invariant. □

THEOREM 3.4. *Let θ_0 be the true parameter, and assume that the true distribution $\pi_{\theta_0} \in G_0(\theta_0)$ is translation invariant but not ergodic. Assume that the ergodic measures $P_{\theta_0}^{(\xi)} \in \mathscr{E}_0(\Phi)$, in $G_0(\theta_0)$ satisfy the conditions of Theorem 3.2. Then $\sqrt{|\Lambda|}\,(\hat{\theta}_\Lambda - \theta_0)$ converges in distribution, to a nonnormal law, as $\Lambda \to \mathscr{Z}^d$.*

PROOF. By the ergodic decomposition (2.9) and Fubini's theorem we have

$$E_{\pi_{\theta_0}}\left\{\exp\left[\mathbf{it}\cdot\sqrt{|\Lambda|}\,(\hat{\theta}_\Lambda - \theta_0)\right]\right\} = \int_{\mathscr{E}_0(\Phi)} \rho_{\pi_{\theta_0}}(d\xi)\, E_{P_{\theta_0}^{(\xi)}}\left\{\exp\left[\mathbf{it}\cdot\sqrt{|\Lambda|}\,(\hat{\theta}_\Lambda - \theta_0)\right]\right\}.$$

Using the dominated convergence theorem, and Theorem 3.2, we obtain

$$\begin{aligned}&\lim_{\Lambda \to \mathscr{Z}^d} E_{\pi_{\theta_0}}\left\{\exp\left[\mathbf{it}\cdot\sqrt{|\Lambda|}\,(\hat{\theta}_\Lambda - \theta_0)\right]\right\} \\ &\quad = \int_{\mathscr{E}_0(\Phi)} \rho_{\pi_{\theta_0}}(d\xi) \lim_{\Lambda \to Z^d} E_{P_{\theta_0}^{(\xi)}}\left\{\exp\left[\mathbf{it}\cdot\sqrt{|\Lambda|}\,(\hat{\theta}_\Lambda - \theta_0)\right]\right\} \\ &\quad = \int_{\mathscr{E}_0(\Phi)} \rho_{\pi_{\theta_0}}(d\xi) \exp\left[-\tfrac{1}{2}\mathbf{t}\cdot C_\xi \mathbf{t}\right],\end{aligned}$$

where

$$C_\xi = T_\xi^{-1} B_\xi T_\xi^{-1},$$

$$T_\xi^{(\alpha,\beta)} = E_{P_{\theta_0}^{(\xi)}}\left\{\frac{\partial \mathscr{U}_0^{(\alpha)}}{\partial x_0}\,\frac{\partial \mathscr{U}_0^{(\beta)}}{\partial x_0}\right\},$$

$$B_\xi^{(\alpha,\beta)} = E_{P_{\theta_0}^{(\xi)}}\left\{\mathbf{Y}_0^{(\alpha)} \sum_{j \in V_0} \mathbf{Y}_j^{(\beta)}\right\}.$$

Since $\exp[(-1/2)\mathbf{t}\cdot C_\xi \mathbf{t}]$ is positive-definite, continuous at $\mathbf{t}=0$ and bounded by 1, we conclude (by the dominated convergence theorem) that

$$\int_{\mathscr{E}_0(\Phi)} \rho_{\pi_{\theta_0}}(d\xi)\exp\left[-\tfrac{1}{2}\mathbf{t}\cdot C_\xi \mathbf{t}\right]$$

is also positive-definite and continuous at $\mathbf{t}=0$ and, hence, it is the characteristic function of a probability distribution. □

4. Numerical experiments. In this section we present six numerical experiments with complete data for testing the VE of (2.24), and one experiment with noisy data for testing the VE of (2.35). In the complete data experiments, we also apply, for comparison, the MPL method. In all six experiments the CPU time used by the VE is much less (by a large factor) than the one used by the MPL method, and the results are comparable. A simulation device, given in Section A, is interesting in itself.

The generic class of Gibbs distribution we use in our experiments is defined as follows: Let $e_\alpha \in \mathscr{Z}^2$, $\alpha = 1,\dots,\infty$, be m distinct vectors in $\mathscr{Z}^2$ (here we work on the lattice $\mathscr{Z}^2$ rather than $\mathscr{Z}^d$). Let Λ be a finite window on $\mathscr{Z}^2$ (typically an $M\times M$ square; in our experiments $M=128$). For simplicity we assume periodic boundary conditions (i.e., Λ is a torus). The Gibbs distributions are given by (1.14) with

$$H_\Lambda^{(\theta)}(x) = \tfrac{1}{2}\sum_{\alpha=1}^{m}\beta^{(\alpha)}\sum_{i\in\Lambda}(x_i - x_{i+e_\alpha})^2 + \sum_{i\in\Lambda} p(x_i;\lambda), \tag{4.1}$$

$$\theta = (\beta,\lambda),\ \beta^{(\alpha)}>0,\ \alpha = 1,\dots,m,\ x_i\in\mathbb{R},$$

where $p(x;\lambda)$ is the polynomial (1.12). For a more general version of (4.1), defined at all levels of resolution or scale, with applications to the representation and synthesis of textures, see [2].

The polynomial $p(x;\lambda)$ is allowed to have one or more local and/or global minima. Intuitively, the configurations $x=\{x_i\colon i\in\Lambda\}$ with high probability are those that tend to minimize $H_\Lambda^{(\theta)}(x)$. The first (quadratic) part in (4.1) induces a cooperation between two interacting pixels i and $i+e_\alpha$, $\alpha=1,\dots,m$. That is, the gray-levels x_i and x_{i+e_α} "tend" to have the same value. This common value is dictated by the minima of $p(x;\lambda)$. Thus, in terms of images, the minima of $p(x;\lambda)$ represent the dominant gray-levels in an image. A high probability sample from $\pi_\theta(x)$ will typically contain "homogeneous" regions, each consisting of values at, or near, a minimum of $p(x;\lambda)$. The heights of the local maxima of $p(x;\lambda)$ control how smooth or how sharp are the transitions from one "homogeneous" region to another "homogeneous" region (see [2] for details and image experiments).

The potentials specified by (4.1) are superstable and regular, and all the assumptions of the previous sections are satisfied. Therefore all our results hold.

A. *Experiments with complete data.* Λ is taken to be a 128×128 square, and we use a special case of (4.1), that is,

$$(4.2) \quad H(x) = \tfrac{1}{2} \sum_{\alpha=1}^{6} \beta^{(\alpha)} \sum_{i \in \Lambda} (x_i - x_{i+e_\alpha})^2 + \sum_{i \in \Lambda} \left(\lambda x_i^4 - \tfrac{1}{2} A x_i^2 + h x_i\right)$$

with $\lambda > 0$, $\beta^{(1)}, \ldots, \beta^{(6)} > 0$ and

$$e_1 = (1,0), \qquad e_2 = (0,1), \qquad e_3 = \frac{1}{\sqrt{2}}(1,1),$$

$$e_4 = \frac{1}{\sqrt{2}}(1,-1), \qquad e_5 = (2,0), \qquad e_6 = (0,2).$$

We assume periodic boundary conditions. If $h = 0$, $\beta^{(1)} = \beta^{(2)} = \beta^{(5)} = \beta^{(6)} \equiv \beta$ and $\beta^{(3)} = \beta^{(4)} = 0$, then with β fixed and $A^2/16\lambda$ very large, a high probability configuration $x = \{x_i : i \in \Lambda\}$ will have components x_i with values near or at $(A/4\lambda)^{1/2}$. Note that $(A/4\lambda)^{1/2}$ is the location of the two wells of the polynomial $\lambda x^4 - 2^{-1}Ax^2$, and $A^2/16\lambda$ is the height (from the global minima) of its local maximum. In this sense, the model behaves like the binary Ising model; in fact there is (see [40], Chapter IX) a much deeper relation between the two models.

First, we simulated the Gibbs distribution associated with (4.2) for six different sets of values of the parameters $\lambda, A, h, \beta^{(1)}, \ldots, \beta^{(6)}$ (see below for method of simulation), and estimated the VEs and MPL estimation. Tables 1–6 show the parameter values (simulation) used in the simulation, the variational estimators (VE) given by (2.24) and the maximum pseudo-likelihood (MPL) estimators. In each experiment, the parameters were estimated from a single realization. The MPL equations were solved using Newton's

TABLE 1

	Parameters								
	λ	A	h	β^1	β^2	β^3	β^4	β^5	β^6
Simulation	50	50	0	10	0.1	0	0	10	0.1
VE	51.146	51.164	0.104	9.920	−0.614	0.0040	0.738	9.071	0.192
MPLE	50.629	50.599	0.0666	9.411	0.192	−0.611	0.541	9.541	0.128

TABLE 2

	Parameters								
	λ	A	h	β^1	β^2	β^3	β^4	β^5	β^6
Simulation	3.364	3.364	0	5.8	5.8	0	0	0	0
VE	3.453	3.256	−0.0198	5.966	6.025	−0.346	−0.125	−0.0579	0.0549
MPLE	3.559	3.428	−0.0134	6.036	5.994	−0.298	−0.155	−0.0701	0.0347

TABLE 3

	Parameters								
	λ	A	h	β^1	β^2	β^3	β^4	β^5	β^6
Simulation	3.364	3.364	0	20	5.8	0	0	0	0
VE	3.650	3.583	0.0247	20.612	5.059	0.131	0.628	−0.517	0.0491
MPLE	3.730	3.687	0.0050	20.489	5.294	0.0183	0.542	−0.388	0.0525

TABLE 4

	Parameters								
	λ	A	h	β^1	β^2	β^3	β^4	β^5	β^6
Simulation	10	10	0	5.8	5.8	0	0	0	0
VE	9.444	9.349	0.0087	6.104	5.874	0.0152	−0.104	−0.0440	−0.153
MPLE	9.683	9.660	0.0052	6.069	5.906	0.0567	−0.144	−0.049	−0.106

TABLE 5

	Parameters								
	λ	A	h	β^1	β^2	β^3	β^4	β^5	β^6
Simulation	100	100	0	10	10	0	0	0	0
VE	103.10	102.80	0.0198	10.493	11.259	−0.545	0.202	0.538	−1.034
MPLE	102.97	102.73	0.0484	10.749	10.900	−0.082	−0.223	0.325	−0.570

TABLE 6

	Parameters								
	λ	A	h	β^1	β^2	β^3	β^4	β^5	β^6
Simulation	100	100	0	10	10	0	0	10	10
VE	99.420	99.572	0.190	10.157	8.701	−1.799	2.472	11.343	9.224
MPLE	98.782	99.035	0.176	10.220	9.674	−1.506	2.257	10.162	9.488

method. The iterations in Newton's algorithm were terminated when the Euclidean norm of the vector that results from the difference between the empirical and the theoretical expectation vectors is smaller than a tolerance error which was set to 10^{-10}. The number of iterations depends, of course, on the initial "guess". The MPL values given in Tables 1–6 were obtained by using as initial "guess" the variational estimators (VE). With these initial values, changes in Newton's algorithm stopped after two iterations for the

second and third experiments (Tables 2 and 3), and after three iterations in the other experiments (Tables 1, 4, 5 and 6).

On an IBM 3090, the CPU time used by the VE was of the order of 1.93 seconds. Each iteration in Newton's algorithm for MPL took about 300 seconds. This shows the computational superiority of the VM over the MPL method. Tables 1–6 show that the two methods give comparable results.

Next we mention briefly three methods that we have used for simulating Gibbs distributions with single pixel random variables $x_i \in \mathbb{R}$: (a) The Langevin equation [19, 15]; (b) The Metropolis algorithm [21, 30, 13]; and (c) The Gibbs sampler [13]. Simulation via the Langevin equation is in general slow, but it has the advantage that it can be implemented easily on parallel (vector) computer architectures. The Metropolis algorithm is the most suitable for Gibbs distributions with continuous random variables. The following implementation makes the Metropolis algorithm especially efficient: Suppose that at the tth sweep ($t = 1, 2, \dots$) we want to update the current value $x_i(t-1)$ at pixel i (given the values at all other pixels). We draw a sample y from a Gaussian distribution on $\mathbb{R}$ with mean $x_i(t-1)$ and variance $\sigma(t)$. Then the new value $x_i(t)$ is chosen to be y or $x_i(t-1)$ according to the Metropolis rule. The covariance $\sigma(t)$ is chosen to decrease monotonically with t. We have not characterized the best schedule for $\sigma(t)$, but in applications the following schedules perform reasonably well:

$$\sigma(t) = \frac{C_0}{1 + \ln t}, \qquad t = 1, 2, \dots$$

and

$$\sigma(t) = \frac{\tilde{C}_0}{t}, \qquad t = 1, 2, \dots$$

with empirically chosen constants C_0 and $\tilde{C}_0$.

The Gibbs sampler is more appropriate for Gibbs distributions with discrete random variables. It can be applied to continuous random variables after discretization ("quantization"), but it is in general expensive. However the data used for the estimations in Tables 1–6 were generated via the Gibbs sampler. The special structure of (4.2) was used to make the Gibbs sampler computationally efficient and highly accurate (in the sense of generating good "typical" samples of the distributions). For example, because of the term λx_i^4, $\lambda > 0$, the contribution to the partition function [see (2.15a)]

$$\int_{\mathbb{R}} \exp\left[-H_i(x_i, x(\mathcal{N}_i))\right] dx_i$$

from large values of $|x_i|$ is small. Thus we chose an interval $[-a, a]$ containing the minima of $\lambda x_i^4 - (A/2)x_i^2 - hx_i$ and subdivided it using a mesh of length $1/32$. The samples used in the estimation of Tables 1–6 were generated after 300 sweeps. We also estimated the same parameters using the sample generated after 1000 sweeps, and the results were approximately the same.

B. *Experiment with noisy data.* We tested the accuracy of the VE provided by (2.35) with a simple experiment: The unobserved process is governed by the exponential family

$$\pi(X) = C(\lambda, A)\exp\{-\lambda X^4 + \tfrac{1}{2}AX^2\}, \qquad X \in \mathbb{R}.$$

We generated 10000 iid samples with $\lambda = 10$, $A = 10$. Then we added a Gaussian noise with mean zero and variance $\sigma^2 = 0.01$. The parameters λ and A were estimated by solving (2.35) via Newton's method. The estimated parameters are $\hat{\lambda} = 9.8285$ and $\hat{A} = 9.8418$. We did also experiments with smaller and higher noise. As expected, the smaller the noise the better the estimated parameters. The accuracy of the estimation was unsatisfactory for large noise variance σ^2.

REFERENCES

[1] ACKLEY, D. H., HINTON, G. E. and SEJNOWSKI, T. J. (1985). A learning algorithm for Boltzman machines. *Cognitive Sciences* **9** 147–169.

[2] ALMEIDA, M. P. (1989). Statistical inference for MRF with unbounded continuous spins and applications to texture representation. Ph.D. dissertation, Div. Appl. Math., Brown Univ.

[3] BESAG, J. (1977). Statistical analysis of non-lattice data. *The Statistician* 179–195.

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

[5] COBB, L., KOPPSTEIN, P. and CHEN, N. H. (1983). Estimation and moment recursion relations for multimodal distributions of the exponential family. *J. Amer. Statist. Assoc.* **78** 124–130.

[6] COMETS, F. and GIDAS, B. (1991). Asymptotics of maximum likelihood estimators for the Curie–Weiss model. *Ann. Statist.* **19** 557–578.

[7] COMETS, F. and GIDAS, B. (1992). Parameter estimation for Gibbs distributions from partially observed data. *Ann. Appl. Probab.* **2** 142–170.

[8] 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.

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

[10] FÖLLMER, H. (1975). Phase transition and the Martin boundary. *Seminaire de Probabilités IX. Lectures Notes in Math.* **465** 305–317. Springer, Berlin.

[11] GEMAN, D., GEMAN, S., GRAFFIGNE, C. and DONG, P. (1990). Boundary detection by constraint optimization. *IEEE Trans. PAMI* **12** 609–628.

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

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

[14] GEMAN, S. and GRAFFIGNE, C. (1987). Markov random field image models and their applications to computer vision. In *Proceedings of the International Congress of Mathematics 1986* (A. M. Gleason, ed.) Amer. Math. Soc., Providence, RI.

[15] GEMAN, S. and HWANG, C. R. (1986). Diffusions for global optimization. *SIAM J. Control Optim.* **24** 1031–1043.

[16] GEMAN, S. and MCCLURE, D. E. (1985). Bayesian image analysis: An application to single photon emission tomography. In *Proc. Amer. Statist. Assoc., Statistical Computing Section*. Amer. Statist. Assoc., Alexandria, VA.

[17] GIDAS, B. (1989). A renormalization group approach to image processing problems. *IEEE Trans. PAMI* **11** 164–179.

[18] GIDAS, B. (1987). Consistency of maximum likelihood and pseudo-likelihood estimators for Gibbs distributions. In *Proceedings of the Workshop on Stochastic Differential Systems with Applications in Electrical/Computer Engineering, Control Theory, and Operations Research*. Springer, Berlin.

[19] GIDAS, B. (1985). Global optimization via the Langevin equation. In *Proceedings of the 24th Conference on Decision and Control* 774–778.

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

[21] GIDAS, B. (1985). Nonstationary Markov chains and convergence of the annealing algorithm. *J. Statist. Phys.* **39** 73–131.

[22] HINTON, G. E. and SEJNOWSKI, T. J. (1983). Optimal perceptual inference. In *Proc. IEEE Conf. Comp. Vision Pattern Recognition*. IEEE, New York.

[23] KÜNSCH, H. (1981). Almost sure entropy and the variational principle for random fields with unbounded state space. *Z. Wahrsch. Verw. Gebiete* **58** 69–85.

[24] LANGE, K. (1991). Convergence of EM image reconstruction algorithms with Gibbs smoothing. Preprint, Dept. Biostatistics, UCLA School of Medicine.

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

[26] LEBOWITZ, J. and PRESUTTI, E. (1980). Statistical mechanics of systems of unbounded spins (erratum). *Comm. Math. Phys.* **78** 151.

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

[28] LEHMANN, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.

[29] LIPPMAN, A. (1986). A maximum entropy method for expert system construction Ph.D. dissertation, Div. Appl. Math., Brown Univ.

[30] METROPOLIS, N., ROSENBLUTH, A., ROSENBLUTH, M., TELLER, A. and TELLER, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21** 1087–1091.

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

[32] PIRLOT, M. (1980). A strong variational principle for continuous spin systems. *J. Appl. Probab.* **17** 47–58.

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

[34] RABINER, L. R. (1988). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *IEEE Proceedings*.

[35] RIPLEY, B. D. (1988). *Statistical Inference for Spatial Processes*. Cambridge Univ. Press.

[36] RUELLE, D. (1970). Superstable interactions in classical statistical mechanics. *Comm. Math. Phys.* **18** 127–159.

[37] RUELLE, D. (1976). Probability estimates for continuous spin systems. *Comm. Math. Phys.* **50** 189–194.

[38] RUELLE, D. (1978). *Thermodynamic Formalism*. Addison–Wesley, Reading, MA.

[39] SILVERMAN B. W., JONES, M. C., WILSON, J. D. and NYCHKA, D. W. (1990). A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography. *J. Roy. Statist. Soc. Ser. B* **52** 271–324.

[40] SIMON, B. (1984). *The $P(\phi)_2$ Euclidean (Quantum) Field Theory*. Princeton Univ. Press.

[41] YOUNES, L. (1988). Problémes d'estimation parametrique par les champs de Gibbs Markoviens. Application au traitment d'image. Thesis, Univesidad Paris-Sud, Orsay.

[42] YOUNES, L. (1988). Estimation and annealing for Gibbs fields. *Ann. Inst. H. Poincaré* **24** 269–294.

DIVISION OF APPLIED MATHEMATICS
BROWN UNIVERSITY
PROVIDENCE, RHODE ISLAND 02912