The Annals of Applied Probability
1995, Vol. 5, No. 3, 577–602

HIDDEN MARKOV RANDOM FIELDS[1]

BY HANS KÜNSCH, STUART GEMAN AND ATHANASIOS KEHAGIAS

ETH Zentrum, Brown University and Brown University

A noninvertible function of a first-order Markov process or of a nearest-neighbor Markov random field is called a hidden Markov model. Hidden Markov models are generally not Markovian. In fact, they may have complex and long range interactions, which is largely the reason for their utility. Applications include signal and image processing, speech recognition and biological modeling. We show that hidden Markov models are dense among essentially all finite-state discrete-time stationary processes and finite-state lattice-based stationary random fields. This leads to a nearly universal parameterization of stationary processes and stationary random fields, and to a consistent nonparametric estimator. We show the results of attempts to fit simple speech and texture patterns.

1. Introduction. If $X = X_1, X_2, \ldots$ is a Markov process and $Y = Y_1, Y_2, \ldots$ is a deterministic or stochastic function of X, then Y is called a *hidden Markov model* (HMM), or sometimes a *hidden Markov process*. Usually, the dependency of Y_t on X is more-or-less local, as when $Y_t = f(X_t)$ for some function f or $Y_t = g(X_t, X_{t+1}, \eta_t)$ for some function g and an iid process $\{\eta_t\}$, independent of X. In any case, Y itself is generally not Markov, and may in fact have a complicated dependency structure. Nevertheless, the *conditional* distribution of X given Y may remain simple, as in the above two examples where X given Y is still first-order Markov. The combination of a rich marginal structure for Y and a simple posterior structure for X makes hidden Markov processes a common modeling tool.

EXAMPLE 1. *Filtering* (cf. [34]). Although the general (nonlinear) filter problem falls within this framework, let us specialize to the linear case: X (known as the *state* process) is not only Markov, but satisfies a simple linear (stochastic) difference equation

$$X_{t+1} = aX_t + \omega_t,$$

where $\{\omega_t\}$ is iid. The *observation* process Y is a HMM, linearly related to X, as in

$$Y_t = bX_t + \omega_t',$$

Received December 1993; revised November 1994.
[1] Supported by Army Research Office Contract DAAL03-92-G-0115 to the Center for Intelligent Control Systems, NSF Grant DMS-88-13699 and Office of Naval Research Contract N00014-91-J-1021.
AMS 1991 *subject classifications*. Primary 60G60; secondary 62M05.
Key words and phrases. Hidden Markov models, Markov random fields, speech models, textures.

where $\{\omega_t'\}$ is another iid noise process, independent of $\{\omega_t\}$. The object is to estimate the state X_t from the observations $\{Y_s\}$, $s \in [0, T]$. This is termed *smoothing* if $0 \leq t < T$, *filtering* if $t = T$ and *prediction* if $t > T$. In any case, the fact that X given Y is still Markov is central to obtaining practical estimation formulas. Beyond this, linearity is exploited to derive efficient recursive estimators (e.g., the Kalman filter) for a host of "on-line" applications in tracking and control.

EXAMPLE 2. *Speech recognition* (see, e.g., [1] and [43]). Here X is a Markov chain with finite (but very large) state space. In principle, the state of X_t represents all of the information relevant to predicting utterances of a speaker at times $\tau > t$. In practice, this information is modeled by representing, jointly, the word (and, sometimes, word pair), phoneme and part of the phoneme (e.g., beginning, middle or end) being articulated at time t. The transition matrix for X is built hierarchically, by successively modeling the variations in pronunciation of parts of phonemes, phonemes and words, as well as (some of) the constraints and regularities in word sequences (syntax). Observations are of the acoustic signal, or some transformation or simplification, and are represented by Y. A stochastic model for Y_t given X_t is developed (or estimated more-or-less nonparametrically). The result is a HMM for the observable acoustic signal (or its transformation) Y, and the object is to estimate X (especially the word sequence) given Y. The posterior is Markov, which is fortunate since this simple dependency structure admits dynamic-programming-like computational tools for the calculation (or at least approximation) of an optimal estimator for X, as well as for computing expectations of various sufficient statistics involved in the estimation of the model parameters. This HMM setup, or some of its variations, is the basis for the most successful speech recognition systems.

EXAMPLE 3. *Ion channel kinetics* (see [3], [2], [24] and [36]). Nerve cells can propagate electrical activity without attenuation over long distances. Lossless conduction involves an active process of opening and closing selective membrane ion channels, and thereby exchanging selected ions between inter- and intracellular spaces. Experiments can be devised to measure the changing conductance of one or a small number of channels in response to various chemical or electrical stimuli. These experiments reveal that ion channels typically move through only a few effective states, being, for example, simple "open" or "closed" with essentially no intermediate levels of conductance. The actual molecular basis for these measurable states is more complicated and is often modeled as a Markov process with multiple states. The observable conductance is then a function of this process, through which, for example, certain of the molecular states manifest themselves as an open channel and others as a closed channel. Thus the observable conductance is a HMM. Purported mechanisms for channel kinetics can be tested by using observed channel conductances to infer the structure and transition probabilities of the (hidden) molecular Markov process. In these applications, the time parameter is generally continuous.

EXAMPLE 4. *Amino acid sequence analysis.* Hundreds or thousands of amino acids strung linearly together constitute a protein. Typically, there are only 20 distinct types of amino acids found, but there are of course a very large number of possible *sequences*. The particular sequence of amino acids that constitutes a protein is known as its "primary" structure. The determination of primary structure is known as sequencing, a process that has been increasingly automated; the result is a large existing data bank of primary structures. The *function* of a protein is largely determined by the folded three-dimensional (or "tertiary") structure that the amino acid chain assumes in vivo. Tertiary structure can sometimes be determined by experimental and imaging techniques, but the process is laborious and the number of sequenced proteins far exceeds the number of proteins with known tertiary structure. Hence, a fundamental problem in biology is the prediction of tertiary structure from primary structure.

One general approach is to search through sequences with known tertiary structures in order to find a "good match" to a sequence with unknown tertiary structure. Similar sequences tend to have similar structure, and in fact there are broad categories of structure that most proteins (or portions thereof) fall into. In an effort to exploit these structural categories, Krogh, Brown, Mian, Sjölander and Haussler [38] built probabilistic models for amino acid sequences conditional on structural classes. These models are built up from known structure-sequence pairs, and then are used to infer a likely structural class for a novel amino acid sequence. Thus, for example, a stochastic model is built for the sequence of amino acids constituting a typical globin (protein that transports oxygen and carbon dioxide). A new amino acid sequence can be evaluated under the globin model to determine its fit, and thereby to predict whether or not it will exhibit a globin-like tertiary structure. Preliminary tests have been highly successful.

The actual models constructed by Krogh, Brown, Mian, Sjölander and Haussler are HMM's with the amino acids constituting the observables and a Markov process, with carefully constructed state space and restricted transitions, constituting the hidden process. (A very similar approach is taken by Churchill [16] in constructing HMM's for the sequence of bases constituting a DNA molecule.) Transition probabilities are estimated from existing data bases, as are state-dependent distributions on the 20 available amino acids. Here again the conditional Markov structure of the unobserved (in fact, virtual) process is heavily exploited to develop computationally feasible estimation and inference algorithms (involving various dynamic-programming-like procedures).

EXAMPLE 5. *Texture models.* This is just a proposal, but it serves to introduce a generalization that will be a primary focus of our theoretical development. Consider a digitized image of a textured pattern such as cloth, wood or sand. The image can be thought of as a realization of a stochastic process $\{Y_t\}$, $t \in \Lambda = \{(i, j)\colon 1 \le i \le N, 1 \le j \le M\}$, where $N = M = 512$, for example, and Y_t is the grey level observed at picture element (or pixel) t. Many authors (e.g., [19], [33] and [21]) have proposed modeling $\{Y_t\}$, condi-

tioned on the texture type and the imaging parameters (distance to camera, orientation, discretization, etc.), as a Markov random field. Since there is usually an organization to the texture that essentially rules out nearest-neighbor models, this approach demands that one either pick, more-or-less arbitrarily, a neighborhood structure or attempt to estimate the neighborhood structure. In either case, there is then the requirement of choosing (or estimating) parameters that determine the associated clique functionals.

A different approach to obtaining the necessary structure would be to employ a *hidden* Markov random field, using a simple nearest-neighbor process for the underlying Markov structure. Thus $Y_t = f(X_t)$, f a fixed "hiding function," where X_t is a nearest-neighbor Markov random field. As in the one-dimensional examples discussed previously, Y will not generally be Markov, although the conditional distribution on X, given Y, is still a nearest-neighbor Markov random field. Is it possible to introduce sufficiently rich structure into the Y process to capture the regularity/variability of real textures through this mechanism? We will return to this shortly.

The last example, especially, raises the issue of generality: How general is the class of processes that can be well approximated by a hidden Markov model? To be concrete, we shall restrict ourselves to nearest-neighbor processes (which is to say, first-order Markov when working in one dimension) and we will only allow instantaneous and deterministic "hiding" functions: $Y_t = f(X_t)$. [In one dimension, many variations are popular: Y_t might depend, randomly or deterministically, on X_t or, simultaneously, on X_t and X_{t-1}. Restricting to *finite* state spaces, it is not difficult to show that these four classes are *equivalent*, in the sense that the set of achievable distributions, for the observable process Y, is identical in each case (see [5] and [35]). One constructs an explicit distribution-preserving transformation from a HMM of one type to a HMM of another type.] Furthermore, X_t (and hence also Y_t) will always have finite state space. So, for example, consider a stationary process $Z_t \in \{0, 1\}$, $t = 1, 2, \ldots$, which we shall try to model (or "fit") with a HMM of the form $Y_t = f(X_t)$, where X_t is first-order Markov, $X_t \in \{0, 1, \ldots, N\}$, f: $\{0, 1, \ldots, N\} \to \{0, 1\}$. By varying N, f and the transition probability matrix for X, how close can we get (how similar to Z can we make Y)?

The answer depends very much on the measure of similarity. Ornstein and Weiss [40], for example, study related questions under a strong notion of similarity: Given two discrete-state stationary processes Y and Z, $d(Y, Z) \leq \varepsilon$ if there exists a stationary process $\Psi = \{\Psi_t\} = \{(Y'_t, Z'_t)\}$ such that:

1. Y' and Z' have the same distributions as Y and Z, respectively.
2. $P(Y'_1 \neq Z'_1) \leq \varepsilon$.

The Ornstein–Weiss *distance*, d, between Y and Z is the *infimum* over all such ε. The results of Ornstein and Weiss indicate that the class of Z which can be arbitrarily well approximated by HMM's Y, relative to d, is highly restricted.

On the other hand, in terms of weak convergence, every stationary Z is the limit of a sequence of HMM's: There exist X^n, first-order Markov on $\{0, 1, \ldots, N_n\}$, f^n: $\{0, 1, \ldots, N_n\} \to \{0, 1\}$, such that $Y^n = f^n(X^n)$ converges weakly to Z as $n \to \infty$ [i.e., for every m, the distribution of $(Z_1, \ldots, Z_m)$ is the limit of the distributions of $(Y_1^n, \ldots, Y_m^n)$]. This is fairly easy. Basically the idea is to define X^n taking values in $\{0, 1\}^n$ ($X_t^n \in \{0, 1\}^n$ for each $t = 1, 2, \ldots$) in such a way that X_{t+1}^n, given X_t^n, has the same distribution as $(Z_{t-n+2}, Z_{t-n+3}, \ldots, Z_{t+1})$ given $(Z_{t-n+1}, Z_{t-n+2}, \ldots, Z_t)$. Then Y_t^n is just the last component of X_t^n (see [40] and [35], and see [41] for versions of this for continuous-valued processes).

The issue of approximating stationary processes by weak limits of HMM's is more complicated in higher dimensions. Let $S = \mathcal{Z}^d$ be the d-dimensional (discrete) square lattice. Let $Z = \{Z_t\}_{t \in S}$ be stationary with finite-state space E ($Z_t \in E$, $\forall t \in S$). The process $X = \{X_t\}_{t \in S}$ is a nearest-neighbor Markov random field (MRF) if the distribution of X_t given $\{X_s\}_{s \neq t, s \in S}$ is the same as the distribution of X_t given $\{X_s\}_{s \in N_t}$, where N_t is the set of $2 \cdot d$ nearest neighbors of t (see [37]). When $d = 1$, this is equivalent to the usual first-order Markov property. Given stationary Z, can we choose an N, an X and an f, such that X is a nearest-neighbor MRF with values in $\{0, 1, \ldots, N\}$, f: $\{0, 1, \ldots, N\} \to E$, and the process Y defined by $Y_t = f(X_t)$ approximates Z? As we shall see shortly (Section 2), there always exists a sequence of these hidden nearest-neighbor Markov random fields that converges weakly to Z. [We use a similar idea as for one dimension: We choose as X_t^n the vector with components Z_s, where s belongs to a block of pixels of size n around t. Actually, we will insist that our hidden process X^n be Gibbs (see Section 2) in addition to being Markov. This entails a modification to enforce strict positivity of the conditional probabilities for X_t^n given $\{X_s^n\}_{s \neq t}$.]

Given a stationary process $Z = \{Z_t\}_{t \in S}$, $S = \mathcal{Z}^d$, taking values in E ($|E| < \infty$), one way to actually build a model for Z would be to try to exploit the above-mentioned result about the (weak) density of HMM's: Search for an N, a nearest-neighbor process $X = \{X_t\}_{t \in S}$ and a function f: $\{0, 1, \ldots, N\} \to E$ such that $Y = \{Y_t\}_{t \in S}$, $Y_t = f(X_t)$, has distribution similar to Z. Actually, f can be fixed, a priori. For example, if $E = \{0, 1, \ldots, M - 1\}$ and $f(x) = x \bmod M$, then the collection of HMM's $Y_t = f(X_t)$, where X_t is a finite state ($\{0, 1, \ldots, N\}$, for some N) nearest-neighbor Markov random field on S, is weakly dense among all stationary Z (with state space $\{0, 1, \ldots, M - 1\}$) on S. Therefore, the construction of a model of this type amounts to choosing a suitable N and an associated process X. If $d = 1$, then X is determined by a transition probability matrix P, which requires specifying approximately N^2 parameters. If $d > 1$, then we can represent X as a Gibbs distribution (see Section 2), which will involve one pair–clique function for each dimension—roughly $d \cdot N^2$ parameters.

In Section 3 we address the problem of modeling Z by *estimating* these parameters via maximum likelihood (ML) or a closely related methodology. We establish a kind of consistency result: Imagine that we are given a sequence of partial observations from a single realization of Z, of the form

$\{Z_t\}_{t \in V_n}$, where $\{V_n\}_{n=1}^{\infty}$ is a sequence of increasing sublattices in $\mathscr{Z}^d$. The number N of states in the hidden process X amounts to a *regularization* or *smoothing* parameter and, as is usual in nonparametric estimation, it will be necessary to relax the smoothing constraint as we accommodate more observations: $N = N_n \uparrow \infty$. We will present conditions under which a maximum likelihood (or closely related) choice of the parameters of the X-process, under the hidden model $Y_t = f(X_t)$, guarantees a consistent estimation of Z, provided $N_n \uparrow \infty$ sufficiently slowly. Convergence is of a relative entropy (between Z and Y) and is almost sure with respect to the distribution of Z. Unfortunately, we can offer no practical recipes for choosing N_n or for actually calculating (global) ML estimates. Nevertheless, we performed some estimation experiments, fixing $N_n = N$ and $V_n = V$, involving acoustic signals from speech and simple binary textures; these are presented in Section 3 as well.

Related work. We have already cited a few related papers. Additionally, several authors have addressed the problem of *identifiability*: Given an HMM Y, describe the (generally large) class of Markov processes X that could, through a suitable hiding function f, generate the distribution of Y. Blackwell and Koopmans [11] seem to have been the first to address the problem. Their results were improved upon by Gilbert [29]. More recently, Itô, Amari and Kobayashi [31] obtained an essentially complete solution. Another related line of research has been the attempt to characterize, in terms of distributional properties, processes Y that are *exactly* functions of Markov chains. Dharmadhikari [20] gave some sufficient conditions and Fredkin and Rice [23] gave some (rather severe and surprising) necessary conditions. A complete algebraic characterization is known, but it is not very manageable —see Chapter III of Rosenblatt [44]. Berbee and Bradley [7], [12] have constructed examples that show that even very rapidly mixing processes need not be HMM's. Brockett's calculations [13] indicate that good approximations of a stationary process by a HMM may require very large state spaces for the underlying Markov process, particularly when the stationary process has a nearly periodic covariance.

Concerning *estimation*, Baum and Petrie [5] established consistency of maximum likelihood estimation of a HMM when the state space of the (hidden) Markov process is known, and more recently Bickel and Ritov [9] extended these results to include information about asymptotic distributions. The problem is, of course, harder when the hidden process (again, with known state space) is a Markov random *field*, but there has been progress here as well; see Comets and Gidas [17] and Frigessi and Piccioni [25]. The issue of how actually to compute maximum likelihood parameter estimators, both for hidden Markov processes and hidden Markov random fields, is discussed by Qian and Titterington [42], who suggest several variations on the EM algorithm [4, 6, 18], and Younes [45], who derives a stochastic gradient ascent algorithm. Finally, we mention the results of Ji [32], who studies nonparametric estimation of certain Gibbs fields. These are related to

our estimation results, since our results amount to a recipe for nonparametrically estimating essentially arbitrary (stationary) random fields (see Section 3), although, unlike Ji, we give no information about rates of convergence.

2. Approximation.

2.1. *Notation and preliminaries.* As in Section 1, S will represent $\mathcal{Z}^d$, the d-dimensional discrete square lattice. Given any finite set E (such as the state space for either the hidden or observable process), the corresponding "configuration space" is $\Omega = E^S = \{x = \{x_t\}_{t \in S}\colon\ x_t \in E\ \forall\, t \in S\}$. The topology on Ω is, as usual, the product topology arising from the discrete topology on E. Similarly, if E' is another finite set, then $\Omega' = E'^S$, again with the product topology. Finally, if $V \subseteq S$ and $x \in \Omega$, then $x_V = \{x_t\}_{t \in V}$.

Gibbs measures (on Ω) are special cases of Markov random fields. We will show that the class of hidden finite-state first-order stationary Gibbs measures is weakly dense among finite-state stationary processes. In particular, this implies the result announced in Section 1, since first-order Gibbs measures are nearest-neighbor Markov random fields.

Gibbs measures arise from *potentials*. For our purposes we will use only shift-invariant and summable potentials. By this we mean a collection of functions $\Phi = \{\Phi_V\}_{V \subset S, V \text{ finite}}$, such that:

1. $\Phi_V\colon\ E^V \to R$.
2. $\Phi_{V+t} = \Phi_V \quad \forall\, t \in S$. More precisely, $\Phi_{V+t} \circ \tau_t = \Phi_V$, where $\tau_t\colon\ \Omega \to \Omega$ is the shift operator $(\tau_t x)_s = x_{s-t}$.
3. $\Sigma_{V \ni 0} \sup_{x \in \Omega} |\Phi_V(x_V)| < \infty$.

A Gibbs measure with potential Φ is any probability measure μ on Ω such that, for any finite $V \subset S$ and $x \in \Omega$,

$$\mu[X_V = x_V | X_{V^C} = x_{V^C}] = \frac{1}{Z} \exp\left\{ - \sum_{\substack{W \subset S \\ W \cap V \neq \varnothing}} \Phi_W(x_W) \right\},$$

where $V^C = S \setminus V$ and Z (which depends on V and x) normalizes the conditional distribution:

$$Z = \sum_{x_V} \exp\left\{ - \sum_{\substack{W \subset S \\ W \cap V \neq \varnothing}} \Phi_W(x_W) \right\}.$$

[The random variable X_t, on Ω, is the coordinate map $X_t(x) = x_t$.]

Define $\delta V = \{t \in S \setminus V\colon\ \exists\ \ s \in V,\ \ W \subset S,\ \ |W| < \infty,\ \ni \Phi_W \neq 0$ and $t,\ s \in W\}$, which is the boundary of V under the neighborhood relation induced by Φ. Then for fixed x_V, $\mu[X_V = x_V \mid X_{V^C} = x_{V^C}]$ depends only on $x_{\delta V}$, so μ is an MRF relative to the neighborhood system

$$N_t = \{s \in S\colon\ s \neq t,\ s, t \in W \text{ some } W \subset S, |W| < \infty, \Phi_W \neq 0\}.$$

In particular, if $\Phi_V = 0$ except for $V = \{t\}$ or $V = \{t, s\}$, where s and t are nearest neighbors in $\mathscr{Z}^d$ (so Φ is a "nearest-neighbor potential"), then μ is a nearest-neighbor Markov random field. A "stationary first-order Gibbs measure" is a stationary Gibbs measure with nearest-neighbor potential.

2.2. *Statement of result.* Given E and E' finite, and a function $f\colon E' \to E$, denote by $\bar{f}$ the function from Ω' to Ω defined by

$$\bar{f}(\{x_t\}_{t\in S}) = \{f(x_t)\}_{t\in S}.$$

Given a measure ν and Ω', define $\mu = \nu \circ \bar{f}^{-1}$ on Ω by

$$\mu(A) = \nu(\bar{f}^{-1}(A)).$$

Now fix E and consider the following sets of probability measures:

$$\mathscr{M}_s = \{\mu \text{ on } \Omega\colon\ \mu \text{ stationary}\},$$

$$\mathscr{M}_g(E') = \{\mu \text{ on } \Omega'\colon\ \mu \text{ stationary first-order Gibbs}\},$$

$$\mathscr{M}_h = \Big\{\mu \text{ on } \Omega\colon\ \mu = \nu \circ \bar{f}^{-1}, \text{ for some } E' \text{ finite},\ \nu \in \mathscr{M}_g(E'), f\colon E' \to E\Big\}.$$

Note that $\mathscr{M}_h$ is the set of hidden finite-state first-order stationary Gibbs measures.

THEOREM 2.2.1. *$\mathscr{M}_h$ is weakly dense in $\mathscr{M}_s$.*

REMARK. In one dimension, $X = \{X_t\}_{t\in\{0,\pm 1,\ldots\}}$ is first-order Markov with positive transition probabilities iff X is first-order Gibbs. Hence, by the theorem, $\{Y\colon\ Y_t = f(X_t)$, some f and some X_t (finite-state) first-order Markov with positive transition probabilities$\}$ is weakly dense among finite-state stationary processes. This special case is fairly easy to get (along the lines of the argument outlined in Section 1). Furthermore, in this case there are results about approximation in the sense of *relative entropy*; see [35].

2.3. *Proof.* The idea is essentially this: Gibbs measures are known to be dense in $\mathscr{M}_s$. Any Gibbs measure can be approximated by a Gibbs measure with potential having finite range $\Phi = \{\Phi_V\}$, where $\Phi_V = 0$ whenever diameter$(V) > B$, for some bound B. Finally, *hidden* first-order Gibbs measures approximate Gibbs measures with finite-range potentials.

In general, there is more than one Gibbs measure with a given potential Φ ("phase transition") and even though Φ is shift-invariant, a Gibbs measure with potential Φ need not be stationary. Denote by $\mathscr{G}_g(\Phi)$ the set of all stationary Gibbs measures with potential Φ. Let $\mathscr{U}$ denote the set of potentials for which $\mathscr{G}_s(\Phi)$ is a singleton and let

$$\mathscr{M}_u = \{\mu \text{ on } \Omega\colon\ \{\mu\} = \mathscr{G}_s(\Phi) \text{ for some } \Phi \in \mathscr{U}\}.$$

Then the following is known ([28], 16.40):

PROPOSITION 2.3.1. *$\mathscr{M}_u$ is weakly dense in $\mathscr{M}_s$.*

Hence, it is sufficient to show that $\mathscr{M}_h$ is weakly dense in $\mathscr{M}_u$. The next step is to truncate Φ in order to have a finite range potential: Let

$$\Phi_V^N = \begin{cases} \Phi_V, & \text{if } V \text{ if contained in } C_N + t \text{ for some } t \in S, \\ 0, & \text{otherwise,} \end{cases}$$

where C_N is the cube $\{-N, -N+1, \ldots, N\}^d \subset S$. Then, in light of the following proposition, it will be sufficient to approximate for each N some member of $\mathscr{G}_s(\Phi^N)$ be a sequence in $\mathscr{M}_h$.

PROPOSITION 2.3.2. *Suppose that $\Phi \in \mathscr{U}$ and that we have an arbitrary sequence of potentials (Φ^N) such that $\Sigma_{V \ni 0} \sup_x |\Phi_V(x_V) - \Phi_V^N(x_V)| \to 0$, as $N \to \infty$. Then for any sequence (μ_N) with $\mu_N \in \mathscr{G}_s(\Phi^N)$ there is a subsequence (N_n) such that $\mu_{N_n} \to \mu$ weakly, where μ is the unique element of $\mathscr{G}_s(\Phi)$.*

PROOF. The hypothesis on (Φ^N) implies that the conditional probabilities

$$\Pi_V^N(x_V \mid x_{V^c}) := \frac{1}{Z_V^N} \exp\Big(- \sum_{W \cap V \neq \varnothing} \Phi_W^N(x_W) \Big)$$

converge in the sup-norm to

$$\Pi_V(x_V \mid x_{V^c}) := Z_V^{-1} \exp\Big(- \sum_{W \cap V \neq \varnothing} \Phi_W(x_W) \Big)$$

for any V. Because Ω is compact, we may assume that $\mu_N \to \nu$ weakly for some $\nu \in \mathscr{M}_s$. We have to show that $\nu \in \mathscr{G}_s(\Phi)$. First, we observe that it is enough to show

$$\int f d\nu = \int \Pi_V f d\nu \tag{1}$$

for any V and any f which depends only on x_V [meaning $f(x) = f(x')$ whenever $x_V = x'_V$]; see [28], 1.24. [Here, $\Pi_V f(x)$ is defined as $\Sigma_{\xi \in E^V} \Pi_V(\xi \mid x_{V^c}) f(\xi)$.] So, we need to prove (1).

Because $\Pi_V^N \to \Pi_V$, we have $\sup_x |\Pi_V^N f(x) - \Pi_V f(x)| \to 0$, so

$$\left| \int f d\nu - \int \Pi_V f d\nu \right| \leq \left| \int f d\nu - \int f d\mu_N \right| + \left| \int f d\mu_N - \int \Pi_V^N f d\mu_N \right| + \varepsilon$$

if N is large enough. The first term goes to zero because $\mu_N \to \nu$, and the second term is zero because $\mu_N \in \mathscr{G}(\Phi^N)$. □

Hence the theorem follows from the next proposition.

PROPOSITION 2.3.3. *Suppose Φ is a potential whose range is contained in some C_N. Then there is a sequence $\{\nu_\beta\}$ of stationary first-order Gibbs measures with state space $E' = E^{C_N}$ and a function f: $E' \to E$ such that $\nu_\beta \circ \bar{f}^{-1}$ converges weakly to some $\mu \in \mathscr{G}_s(\Phi)$ as $\beta \to \infty$.*

PROOF. Since N is fixed we write C instead of C_N. If $\{x_t\} \in \Omega = E^S$, then we define $\{y_t\} \in \Omega' = E'^S$ by

$$y_t = x_{C+t}. \tag{2}$$

We index the components of y_t by $r \in C$, rather than choosing an arbitrary enumeration. So the rth component of y_t, $y_{t,r}$, is equal to x_{t+r}. We will use the symbol y always to denote an element of Ω' of the form given by (2). An arbitrary element of Ω' will be written as z or ξ. An equivalent way to express that y is of the form (2) is via the following compatibility constraints:

$$(3) \qquad y_{t,r} = y_{s,r+t-s} \qquad \forall\, r, t, s \text{ with } r \in C \text{ and } r + t - s \in C.$$

Moreover the following apparently weaker form is also equivalent to (3) and (2):

$$(4) \qquad y_{t,r} = y_{s,r+t-s} \qquad \forall\, r, t, s \text{ with } \|t - s\| = 1, r \in C \text{ and } r + t - s \in C.$$

This can be seen by connecting t and s through a chain $t = t_0, t_1, \ldots, t_n = s$ such that $\|t_{i+1} - t_i\| = 1$ and $r + t_{i+1} - t_i \in C$.

We now turn to the definition of the approximating potential Φ^β (which will define a first-order Gibbs measure on Ω'). To motivate this definition, note that $\{y_t\}$ is a sample from a first-order Markov random field if $\{x_t\}$ is a sample from some $\mu \in \mathcal{G}_s(\Phi)$. However, this Markov random field is not Gibbs because of the hard constraints (4). We change these hard constraints into soft ones by introducing potentials with value β for each of the constraints (4) which is violated. By letting β tend to infinity we hope to recover (4) and at the same time to obtain the right distribution on the configurations satisfying (4). We show that this is indeed the case.

We seek to approximate the potential Φ. Without limitation of generality we may assume

$$\Phi_V \equiv 0 \quad \text{if } V \neq t + C \text{ for some } t \in S.$$

Then we define Φ^β as

$$\Phi^\beta_{\{t\}}(z_t) = \Phi_C(z_t),$$

$$\Phi^\beta_{\{t,s\}}(z_t, z_s) = \beta\Psi(z_t, z_s) := \beta \sum_{\substack{r, \\ r \in C \text{ and} \\ r+t-s \in C}} 1_{[z_{t,r} \neq z_{s,r+t-s}]} \quad \text{if } \|t - s\| = 1,$$

$$\Phi^\beta_V = 0, \quad \text{otherwise.}$$

We denote by π^β_V the conditional probabilities associated with Φ^β and choose some $\nu_\beta \in \mathcal{G}_s(\Phi^\beta)$. Because of compactness, we may assume that $\nu_\beta \to \nu$ weakly as $\beta \to \infty$. We have to show that $\nu \circ f^{-1} \in \mathcal{G}_s(\Phi)$, where $f(z_t) = z_{t,0}$. The following lemma is the key.

LEMMA 2.3.1. *For arbitrary V consider the event $A = \{z \in \Omega' \mid \Psi(z_r, z_s) = 0 \ \forall\, t, s \in V, \|t - s\| = 1\}$. Then $\nu(A) = 1$. That is, the compatibility constraints are fulfilled a.s.*

We defer the proof of this lemma to the end and consider the conditional distribution of $z_{0,0}$ given $z_{s,0}$, $s \neq 0$, under ν. We choose $K > N$ arbitrarily (recall that N is the size of the box which contains the potential Φ) and put

$$\Lambda = C_K \setminus C_N,$$
$$\overline{\Lambda} = C_{K+N} \setminus \{0\}.$$

We want to show that

$$\nu\big(z_{0,0} = x_0 \mid z_{s,0} = x_s, s \in \overline{\Lambda}\big) \propto \exp\Big\{-\sum_{t \in C} \Phi_C(x_{C+t})\Big\}. \tag{5}$$

Note that knowing x_s, $s \in \overline{\Lambda}$, is the same as knowing y_s, $s \in \Lambda$. Hence, by Lemma 2.3.1 above,

$$\nu\big(z_{s,0} = x_s, s \in \overline{\Lambda}\big) = \nu(z_s = y_s, s \in \Lambda).$$

Moreover, if this probability is positive, then by weak convergence of ν_β to ν,

$$\begin{aligned}\nu(z_{0,0} = x_0 \mid z_s = y_s, s \in \Lambda) &= \lim_{\beta \to \infty} \nu_\beta(z_{0,0} = x_0 \mid z_s = y_s, s \in \Lambda) \\ &= \sum_{\substack{z_C \\ z_{0,0} = x_0}} \lim_\beta \Pi_C^\beta(z_C \mid y_\Lambda).\end{aligned}$$

But by the definition of Φ^β we have, with $x_0 = z_{0,0}$,

$$\Pi_C^\beta(z_C \mid y_\Lambda) = Z^\beta(y_\Lambda)^{-1} \exp\Bigg\{-\sum_{t \in C} \Phi_{\{t\}}^\beta(z_t) - \sum_{\substack{s,t \in C \\ \|s-t\|=1}} \Phi_{\{s,t\}}^\beta(z_s, z_t) - \sum_{\substack{s \in C \\ t \in \Lambda \\ \|s-t\|=1}} \Phi_{\{s,t\}}^\beta(z_s, y_t)\Bigg\}.$$

As $\beta \to \infty$, this converges to

$$\frac{\exp\{-\Sigma_{t \in C} \Phi_C(x_{C+t})\}}{\Sigma_{\substack{\bar{x}: \bar{x}_s = x_s \\ \forall s \neq 0}} \exp\{-\Sigma_{t \in C} \Phi_C(\bar{x}_{C+t})\}}$$

if $z_t = y_t = x_{C+t}$, $\forall\, t \in C$, and to zero otherwise. Hence we have proven (5). [If $\nu(z_s = y_s, s \in \Lambda) = 0$, we can define the conditional probability such that (5) holds.] Because K is arbitrary, (5) says that $\nu \circ \bar{f}^{-1}$ has the required conditional distribution for $t = 0$, and thus for any t by stationarity. This implies that $\nu \circ \bar{f}^{-1} \in \mathcal{G}_s(\Phi)$ ([28], 1.33). □

It remains to prove Lemma 2.3.1.

PROOF OF LEMMA 2.3.1. For any Λ, define the energy in Λ given the boundary conditions by

$$H_\Lambda^\beta(z) = \sum_{W \cap \Lambda \neq \varnothing} \Phi_W^\beta(z_W).$$

Choose Λ such that $V + 2C = V + C + C \subset \Lambda$. We will show that we can modify an arbitrary configuration $z \notin A$ to one which belongs to A, has the same boundary conditions with respect to Λ and whose energy in Λ is smaller by an amount β. Hence

$$\Pi_\Lambda^\beta[A \mid z_{S\setminus\Lambda}] \to 1 \quad \text{as } \beta \to \infty,$$

uniformly in $z_{S\setminus\Lambda}$. By integrating over the boundary conditions we thus obtain

$$\nu_\beta[A] \to 1 \quad \text{as } \beta \to \infty.$$

The modification mentioned above goes as follows: Let $g\colon \Omega' \to \Omega'$ be defined by

$$g(z)_{s,r} = \begin{cases} z_{s+r,0}, & \text{if } r \in C \text{ and } s + r \in V + C, \\ z_{s,r}, & \text{otherwise.} \end{cases}$$

Then it is clear that $g(z)_s = z_s$ if $s \notin V + 2C$, and thus the boundary condition does not change. Also, it is seen easily that $g(z)$ belongs to A. So let us compare the energies $H_\Lambda^\beta(z)$ and $H_\Lambda^\beta(g(z))$ for $z \notin A$:

$$\sum_{t\in\Lambda} \Phi_{\{t\}}^\beta(g(z)_t) \le \sum_{t\in\Lambda} \Phi_{\{t\}}^\beta(z_t) + |V + 2C|\delta,$$

$$\text{where } \delta = \sup \Phi_C(x_C) - \inf \Phi_C(x_C),$$

because $g(z)_t = z_t$ if $t \notin V + 2C$. Moreover, because $z \notin A$ and $g(z) \in A$,

$$\sum_{\substack{t,s\in V \\ \|t-s\|=1}} \Psi(z_t, z_s) \ge 1, \qquad \sum_{\substack{t,s\in V \\ \|t-s\|=1}} \Psi(g(z)_t, g(z)_s) = 0.$$

Now take a $t \notin V$ and s arbitrary with $\|t - s\| = 1$. If $t + r \in V + C$ and $r + t - s \in C$, then $g(z)_{t,r} = g(z)_{s,r+t-s}$, and if $t + r \notin V + C$ and $r + t - s \in C$, then $g(z)_{t,r} = z_{t,r}$ and $g(z)_{s,r+t-s} = z_{s,r+t-s}$. Hence $\Psi(g(z)_t, g(z)_s) \le \Psi(z_t, z_s)$. Together we obtain

$$H_\Lambda^\beta(g(z)) \le H_\Lambda^\beta(z) - \beta + |V + 2C|\delta.$$

This completes the proof of the lemma. □

3. Estimation. The approximation result of Section 2 suggests modeling stationary processes with hidden nearest-neighbor MRF's, or simply (hidden) first-order Markov processes in the one-dimensional case. A single sample path from an ergodic stationary process should be sufficient to determine the parameters for an approximation of this type, and this is confirmed, roughly speaking, by our consistency results: Given a sequence of observations from a single sample of an ergodic stationary process, we use Grenander's method of sieves [30] to construct a sequence of hidden first-order processes with distributions converging (in the sense of relative entropy) to the stationary process.

Actually, we will need to restrict the class of stationary processes somewhat when working in one dimension, and somewhat more when working in higher dimensions; see Section 3.1.

Two sets of experiments were performed. Data from speech signals and textures were used to fit hidden nearest-neighbor processes, via maximum likelihood, and the resulting models were sampled and compared to the original data; see Section 3.2.

3.1. *Consistency.* There are two theorems, for random processes ($S = \mathcal{Z}$) and random fields ($S = \mathcal{Z}^d$, $d > 1$), respectively. The two proofs follow the same general plan, which we will present, in brief outline, for the one-dimensional ($S = \mathcal{Z}$) case. The full details are available in a technical report; see [27].

We imagine observing a stationary process Z with state space $E = \{0, 1, \ldots, M-1\}$, for some (finite) $M > 1$. Let μ_o be the (unknown) distribution, or law, of Z. Following the notation of Section 2, the process Z is to be approximated by a hidden process $Y = \bar{f}(X)$, where X is nearest-neighbor with state space $E' = \{0, 1, \ldots, N\}$. Henceforth, the hiding function f (and consequently $\bar{f}$ as well) is fixed: $f(x) = x \bmod M$. Specializing to the one-dimensional problem ($S = \mathcal{Z}$), X is first-order Markov, and we will adopt the standard representation in terms of transition probability matrices rather than using potentials and the Gibbs representation. Let

$$\mathcal{M}_N = \Big\{ m = \{m_{ij}\}_{i,j=0}^{N} : \ m \text{ trans. prob. matrix,} \\ \text{and } m_{ij} \geq e^{-N} \ \forall \ 0 \leq i, j \leq N \Big\}.$$

The parameter N will serve as a "regularization" or "smoothing" parameter and will eventually be tied to the number n of observations, $Z_0 = z_0$, $Z_1 = z_1, \ldots, Z_n = z_n$, through an increasing function. For any $m \in \mathcal{M}_N$, denote by μ_m the distribution of the hidden Markov process $Y = \{Y_t\}$, $Y_t = f(X_t)$, where $\{X_t\}$ is the unique stationary Markov process with transition matrix m. The results of Section 2 suggest that μ_o can be approximated by a distribution μ_m, for suitable m and large enough N. Having observed $Z_0 = z_0$, $Z_1 = z_1, \ldots, Z_n = z_n$, we denote by $\mathrm{ML}_{N,n}$ the set of maximum likelihood matrices from within $\mathcal{M}_N$:

$$\mathrm{ML}_{N,n} = \mathrm{ML}_{N,n}(z) \\ = \Big\{ m \in \mathcal{M}_N : \ \mu_m(z_0, z_1, \ldots, z_n) = \sup_{q \in \mathcal{M}_N} \mu_q(z_0, z_1, \ldots, z_n) \Big\}.$$

(In general, $\mathrm{ML}_{N,n}$ has more than one element. In any case, it is never empty: $\mathcal{M}_N$ is compact and μ_q is continuous in q.) Under an additional condition on Z, there exists a sequence (N_n) such that the set of HMM's associated with $\mathrm{ML}_{N_n,n}$ is consistent for μ_o:

THEOREM 3.1.1. *Let $\{Z_t\}_{t=-\infty}^{\infty}$ be a stationary ergodic process with finite state space, $Z_t \in \{0, 1, \ldots, M-1\}$, $M < \infty$, and distribution function μ_o. If $\exists$ $\delta > 0 \ni \mu_o(z_0 \mid z_1, \ldots, z_{-t}) \geq \delta \ \forall \ t, (z_0, \ldots, z_{-t}) \in \{0, 1, \ldots, M-1\}^{t+1}$, then for all $N_n \uparrow \infty$ sufficiently slowly,*

$$\sup_{m \in \mathrm{ML}_{N_n, n}} \int \log \frac{\mu_o(z_0 | z_{-1}, z_{-2}, \ldots)}{\mu_m(z_0 \mid z_{-1}, z_{-2}, \ldots)} \, d\mu_o(z) \to 0 \quad a.s.\ (\mu_o).$$

REMARKS.

1. More precisely, there exists a sequence $N_n \uparrow \infty$ such that the assertion holds for all sequences $N_n' \uparrow \infty$ satisfying $N_n' \leq N_n \ \forall \ n$.
2. Unfortunately, $N_n = N_n(\mu_o)$. Roughly speaking, $\{Z_t\}$ can yield information arbitrarily slowly.
3. There is nothing special about the regularization $m_{ij} \geq e^{-N}$. If instead, $m_{ij} \geq g(N)$, where $g(N) \downarrow 0$, then there will be a relationship between $g(N)$ and N_n such that the *faster* $g(N) \downarrow 0$, the *slower* $N_n \uparrow \infty$, in order to insure consistency.
4. We do not know whether we have convergence also with respect to the weak topology of measures.

The corresponding result for Z on $S = \mathscr{Z}^d$, $d > 1$, is somewhat more complicated, even to state. First, we make the additional assumption that μ_o, the distribution of Z, is in fact a (ergodic and stationary) Gibbs measure (see Section 2). In other words, we shall assume that μ_o is a measure on $\{0, 1, \ldots, M-1\}^S$ satisfying (i) μ_o is stationary and ergodic, and (ii) for every finite $V \subset S$,

$$\mu_o[Z_V = z_V \mid Z_{V^c} = z_{V^c}] \propto \exp\left\{ - \sum_{\substack{W \subset S \\ W \cap V \neq \varnothing}} \Phi_W(z_W) \right\},$$

where $\Phi = \{\Phi_V\}$, $V \subset S$, finite, is a (shift-invariant, summable) *potential*, as defined in Section 2. [Obviously (ii) implies that the conditional distributions are bounded from below as was required already in Theorem 3.1.1. In addition (ii) also implies that the conditional distributions are continuous in z. As a converse, boundedness and continuity of the conditional distributions imply (ii) ([28], 2.30). Furthermore, by Proposition 2.3.1 above we also know that the set of μ_o's satisfying (i) and (ii) is weakly dense in the set of stationary measures since uniqueness of a Gibbs measure implies that it is stationary and ergodic ([28], 5.11 and 14.15).]

We can no longer index the approximating measures by transition probabilities. Instead, we replace $\mathscr{M}_N$ by a set of "regularized" potentials $\mathscr{P}_N$:

$$\mathscr{P}_N = \{\text{nearest neighbor potentials } \Psi = (\Psi_0, \Psi_1, \ldots, \Psi_d) \text{ on } \{0, 1, \ldots, N\} \text{ with bounds } |\Psi_0(k)| \leq N, |\Psi_i(j, k)| \leq N, 1 \leq i \leq d\},$$

with the understanding that Ψ_0: $\{0, 1, \ldots, N\} \to R$ is the one-point potential $\Phi_{\{t\}}$ and Ψ_i: $\{0, 1, \ldots, N\}^2 \to R$ are the pair potentials $\Phi_{\{t,t+e_i\}}$ $(1 \le i \le d)$, where e_i is the vector with $d-1$ zeros and a single one at the ith component. All other potentials Φ_V are identically zero. A Gibbs measure ν with potential $\Psi \in \mathscr{P}_N$ is then defined as in Section 2.1. Evidently, ν is then nearest-neighbor Markov.

For any $\Psi \in \mathscr{P}_N$, let $\mathscr{G}_s(\Psi)$ be the set of stationary Gibbs measures with potential Ψ. The set $\mathscr{G}_s(\Psi)$ is always nonempty, but may contain more than one measure. Therefore, there is a *set* of hidden Gibbs measures associated, through f, with each $\Psi \in \mathscr{P}_N$. We denote this set by $\mathscr{H}(\Psi)$:

$$\mathscr{H}(\Psi) = \{\mathscr{L}(Y)\colon\ Y_t = f(X_t), \mathscr{L}(X) \in \mathscr{G}_s(\Psi)\},$$

where $\mathscr{L}(\cdot)$ is the distribution (or law) of a process.

Now suppose that we observe $Z_V = z_V$, where $\mathscr{L}(Z) = \mu_o$ and $V \subset S$ is finite. The idea is to choose a maximum likelihood potential Ψ from within $\mathscr{P}_N$, in other words to choose $\Psi \in \mathscr{P}_N$ in such a way that the associated hidden Gibbs measures assign maximum probability (likelihood) to z_V. Unfortunately, given a candidate potential $\Psi \in \mathscr{P}_N$, the likelihood of z_V under the hidden Gibbs model associated with Ψ is not necessarily well defined; different elements of $\mathscr{H}(\Psi)$ may assign different likelihoods to z_V. Furthermore, even when $\mathscr{G}_s(\Psi)$ contains only one measure, the actual calculation of the probability of z_V under the associated hidden measure is intractable. For these reasons we will employ the following modification of the likelihood.

Fix, once and for all, a configuration $x \in \Omega = \{0, 1, \ldots, M-1\}^S$. For any $\Psi \in \mathscr{P}_N$, define the (conditional) log-likelihood

$$L_V(\Psi, z_V) = \log\left\{ \sum_{\substack{\xi_V\colon \xi_t \in f^{-1}(z_t) \\ t \in V}} \mu[\,\xi_V \mid x_{V^c}]\right\},$$

where $\mu \in \mathscr{G}_s(\Psi)$. L is well defined: it is independent of which $\mu \in \mathscr{G}_s(\Psi)$ we choose. Furthermore, L depends only on $x_{\delta V}$, where δV is the boundary of V under the nearest-neighbor system in $\mathscr{Z}^d$.

Finally, define $M_{N,V}$ to be the set of maximum likelihood potentials within $\mathscr{P}_N$:

$$M_{N,V} = M_{N,V}(z) = \left\{\Psi \in \mathscr{P}_N\colon\ L_V(\Psi, z_V) = \sup_{\Phi \in \mathscr{P}_N} L_V(\Phi, z_V)\right\},$$

and for any two stationary probability measures μ and ν, define $h(\mu, \nu)$ to be the *specific relative entropy* (see, e.g., [28]):

$$h(\mu, \nu) = \liminf_{V \uparrow S} \frac{1}{|V|} E_\mu\left[\log \frac{\mu[x_V]}{\nu[x_V]}\right].$$

By Jensen's inequality, $h(\mu, \nu) \ge 0$ and $h(\mu, \nu) = 0$ if $\mu = \nu$. Conversely, if ν is Gibbs with summable potential and $h(\mu, \nu) = 0$, then also μ is Gibbs with the same potential as ν ([28], 15.37).

THEOREM 3.1.2. *Let μ_o be an ergodic stationary Gibbs measure on $\{0, 1, \ldots, M-1\}^S$, $S = \mathbb{Z}^d$, and let (V_n) be an increasing sequence of finite subsets of S, such that $S = \bigcup_{n=1}^{\infty} V_n$ and $|\delta V|/|V| \to 0$. For all sequences $N_n \uparrow \infty$ sufficiently slowly,*

$$\sup_{\Psi \in \mathrm{ML}_{N_n, V_n}} \sup_{\mu \in \mathscr{X}(\Psi)} h(\mu_o, \mu) \to 0 \quad a.s.\ (\mu_o).$$

REMARK. For $d = 1$ the claim of Theorem 3.1.2 is the same as of Theorem 3.1.1, that is,

$$h(\mu_o, \mu) = \int \log(\mu_o(z_0 \mid z_{-1}, \ldots)/\mu(z_0 \mid z_{-1}, \ldots))\, d\mu_o(z)$$

for the law μ of any hidden Markov process. This is easily seen by an argument following the Shannon–McMillan–Breiman theorem, since for any such μ,

$$\mu(z_0 \mid z_{-1}, \ldots, z_{-k}) \to \mu(z_0 \mid z_{-1}, \ldots)$$

uniformly in z (cf. [27]).

PROOF OF THEOREM 3.1.2 (Outline). The approach is substantially the same for both consistency results. Let us consider the case of μ_o defined on $\{0, 1, \ldots, M-1\}^{\mathbb{Z}}$ (i.e., dimension 1), and go through a brief outline of the proof. (The details for both the consistency theorems are available through the technical report [27].)

The proof is based upon two lemmas. The first is a kind of uniform law of large numbers for the probabilities μ_m, $m \in \mathscr{M}_n$, reminiscent of the Shannon–McMillan–Breiman theorem (cf. [10]).

LEMMA 3.1.1.

$$\lim_{n \to \infty} \sup_{m \in \mathscr{M}_{N_n}} \left| \frac{1}{n} \log \mu_m(z_0, z_1, \ldots, z_n) - \int \log \mu_m(z_0 \mid z_{-1}, z_{-2}, \ldots)\, d\mu_o(z) \right| = 0 \quad a.s.\ (\mu_o)$$

for all $N_n \uparrow \infty$ sufficiently slowly.

The second lemma insures that there is *some* sequence $m_N \in \mathscr{M}_N$ such that μ_{m_N} approaches μ_o.

LEMMA 3.1.2. *There exists a sequence of matrices $m_N \in \mathscr{M}_N$ such that*

$$\lim_{N \to \infty} \int \log \mu_{m_N}(z_0 \mid z_{-1}, z_{-2}, \ldots)\, d\mu_o(z) = \int \log \mu_o(z_0 \mid z_{-1}, z_{-2}, \ldots)\, d\mu_o(z).$$

(The proof of Lemma 3.1.2 is by construction.)

Now assume that the lemmas are true. By Jensen's inequality,

$$\int \log \mu_m(z_0 \mid z_{-1}, \dots)\, d\mu_o(z) \le \int \log \mu_o(z_0 \mid z_{-1}, \dots)\, d\mu_o(z),$$

for all N and $m \in M_N$, so it is enough to show that

$$\liminf_{n\to\infty} \inf_{m \in \mathrm{ML}_{N_n,n}} \int \log \mu_m(z_0 \mid z_{-1}, \dots)\, d\mu_o(z)$$

$$\ge \int \mu_o(z_0 \mid z_{-1}, \dots)\, d\mu_o(z) \quad \text{a.s.}$$

By application of the lemmas,

$$\liminf_{n\to\infty} \inf_{m \in \mathrm{ML}_{N_n,n}} \int \log \mu_m(z_0 \mid z_{-1}, \dots)\, d\mu_o(z)$$

$$= \liminf_{n\to\infty} \inf_{m \in \mathrm{ML}_{N_n,n}} \left\{ \left(\int \log \mu_m(z_0 \mid z_{-1}, \dots)\, d\mu_o(z) - \frac{1}{n} \log \mu_m(z_0, z_1, \dots, z_n) \right) + \frac{1}{n} \log \mu_m(z_0, z_1, \dots, z_n) \right\}$$

$$\ge \liminf_{n\to\infty} \inf_{m \in \mathrm{ML}_{N_n,n}} \left\{ \frac{1}{n} \log \mu_m(z_0, z_1, \dots, z_n) - \left| \frac{1}{n} \log \mu_m(z_0, z_1, \dots, z_n) - \int \log \mu_m(z_0 \mid z_{-1}, \dots)\, d\mu_o(z) \right| \right\}$$

$$= \liminf_{n\to\infty} \inf_{m \in \mathrm{ML}_{N_n,n}} \frac{1}{n} \log \mu_m(z_0, z_1, \dots, z_n) \quad \text{(a.s., by Lemma 3.1.1)}$$

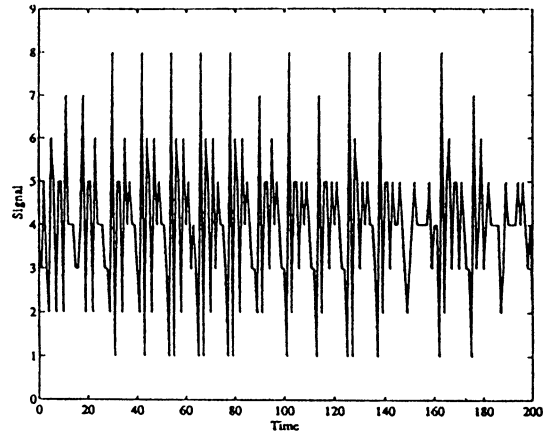
$$\ge \liminf_{n\to\infty} \frac{1}{n} \log \mu_{m_{N_n}}(z_0, z_1, \dots, z_n)$$

$$= \liminf_{n\to\infty} \int \log \mu_{m_{N_n}}(z_0 | z_{-1}, \dots)\, d\mu_o(z) \quad \text{(again, a.s., by Lemma 3.1.1)}$$

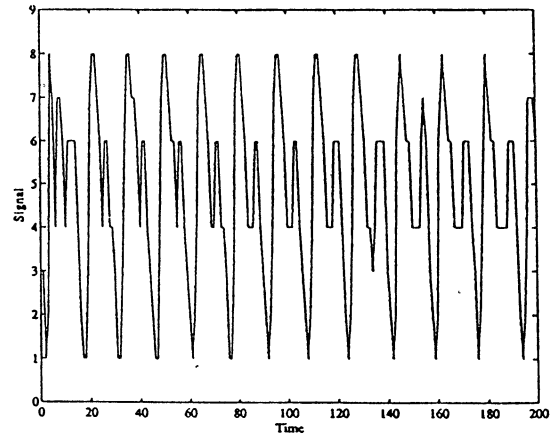
$$= \int \log \mu_o(z_0 \mid z_{-1}, \dots)\, d\mu_o(z) \quad \text{(by Lemma 3.1.2).} \qquad \square$$

3.2. *Experiments.* Consistency is reassuring, but it tells us too little about performance on real (finite) data. We have therefore run some simple experiments in order to assess the "finite sample" promise of the proposed models. There were two kinds of experiments: one-dimensional estimation experiments from speech waveforms and two-dimensional estimation experiments from simple binary textures.

3.2.1. *Speech waveforms.* Segments of two phonemes were extracted from a single utterance of the word "one"; see Figure 1. The acoustic signal was sampled at 10 kHz (one sample every 0.1 ms) and 4096 amplitude levels, although each amplitude was later rounded to one of eight equally spaced values, and the samples themselves were subsampled to one observation at every 0.5 ms. Panel (a) shows a 100-ms segment from the phoneme /a/, which follows the initial /u/ and precedes the final /n/ in the pronunciation of "one." There are 200 data points, each having one of eight values. Panel (b) shows an analogous segment from the final /n/ of the same utterance. The nearly periodic waveforms are characteristic of so-called voiced phonemes, and derive ultimately from more-or-less periodic oscillations of the vocal chords.



(a)



(b)

FIG. 1. (a) 100-*millisecond segment from the phoneme* /a/. (b) 100-*ms segment from the phoneme* /n/.

We treated each signal as a sample $z_1, z_2, \ldots, z_{200}$, from an eight-valued stationary process, which we attempted to fit with a series of hidden Markov models of increasing size. In each case, we employed the "hiding function" $f(x) = 1 + x \bmod 8$, and computed approximate maximum likelihood $N \times N$ transition probability matrices, for $N = 10$, 20, 30, 40, 50 and 60. Maximum likelihood computations were made via the Baum reestimation formula [4, 6], which is an instance of the EM procedure [18]. Estimates were only approximately maximum likelihood since this is an iterative hill-climbing algorithm; it can approach a local maximum and, as a practical matter, it must be terminated short of convergence. We began each run (one run for each value of N) with a randomly generated transition probability matrix and continued until there were only negligible changes in the transition matrix.

The results are most easily judged by viewing samples from the resulting HMM's. Figures 2 (for the /a/ sequence) and 3 (for the /n/ sequence) show random samples from $Y_t = f(X_t^N)$, where $\{X_t^N\}_{t=1}^{200}$ is first-order Markov on $\{0, 1, \ldots, N-1\}$ with the estimated $N \times N$ transition probability matrix, and $X_1^N = 1$ ($N = 10$, 20, 30, 40, 50 and 60). In both sets of experiments, one gets the impression that the fit generally improves with increasing N, although there is the suggestion of some deterioration at $N = 60$. Since there are only 200 (highly correlated) samples, it may be that, at $N = 60$, the familiar problem of over-fitting has been encountered. There may, as well, be computational problems with the iteration procedure, perhaps related to local maxima. In any case, it would be interesting to perform similar experiments with larger data sets; essentially infinite amounts of data are easily available.

It may also be interesting to splice together such signals, as a novel approach to speech synthesis. In this regard, one would need to fit, as well, the nonstationary speech units associated with various consonants. Because we are after a signal of only finite duration, it is not impossible, and perhaps not unreasonable to speculate, that exactly the same models would be effective for fitting consonants.

An obvious alternative approach would be to fit each signal with an Nth-order Markov process. However, even at the modest eight-level discretization used in our experiments, this would involve estimating $7 \cdot 8^N$ parameters, which evidently places a severe restriction on the process order. It may be true, in contrast, that the hidden process provides an efficient coding of the nearly periodic structure by dedicating single or multiple states to positions within the cycle, although we have performed no systematic experiments to test this conjecture.

3.2.2. *Binary textures.* The experiments with two-dimensional processes were more difficult and less successful. We adopted the modest goal of fitting some simple binary textures. These were derived from real textures, borrowed from the well-used Brodatz collection [14], by simply thresholding grey-level pictures. A suitable threshold produces substantial islands of "ones" positioned among a sea of "zeros." The shape and pattern of the islands, of course, depends upon the texture. Figure 4 has two examples: straw and paper. In each, there are $80 \times 60 = 4800$ pixels; "ones" are depicted with dots and "zeros" with stars.

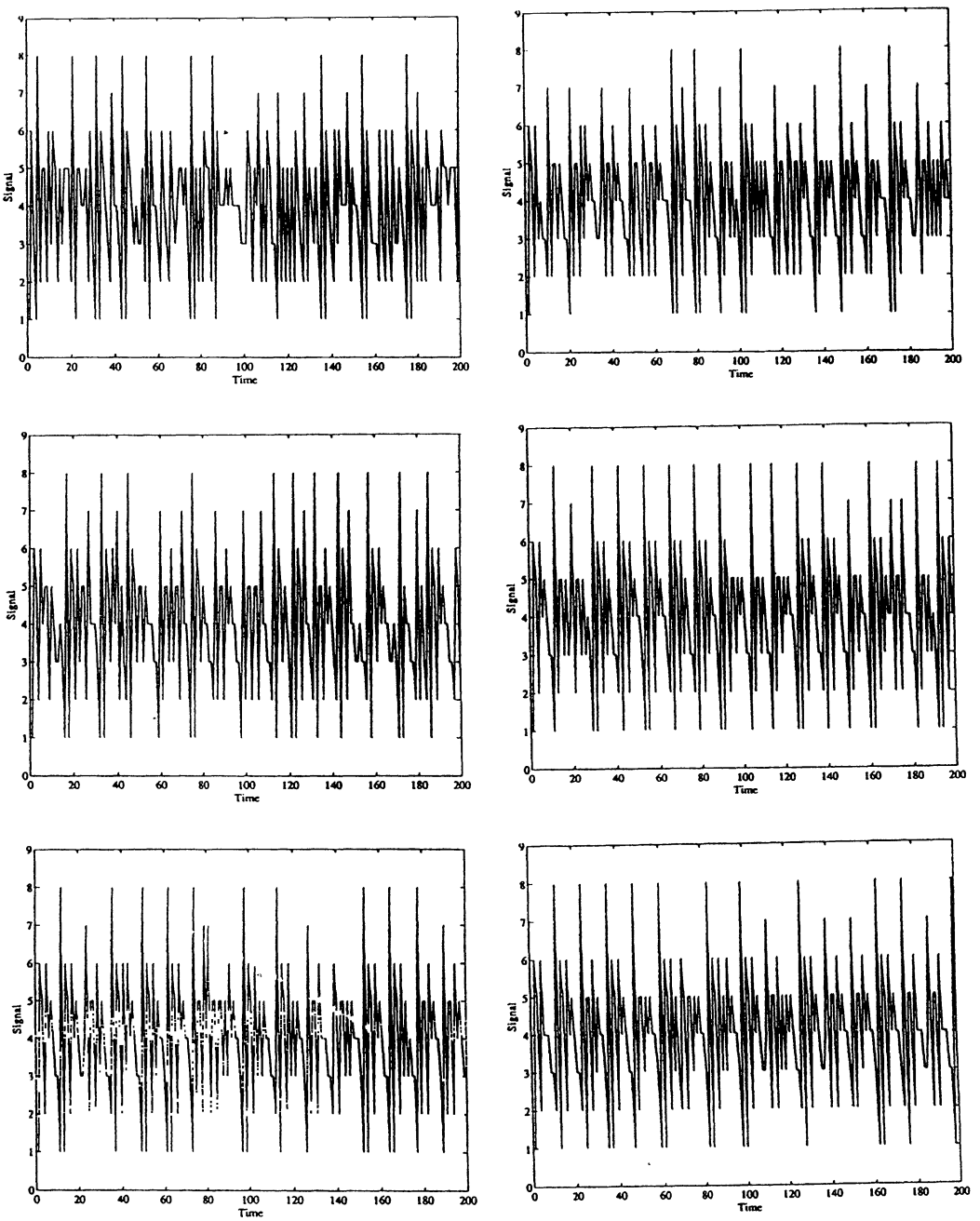


FIG. 2. *HMM's estimated from data in Figure* 1a. *Top to bottom, left-hand side*: 10, 20 *and* 30 *hidden states. Top to bottom, right-hand side*: 40, 50 *and* 60 *hidden states.*

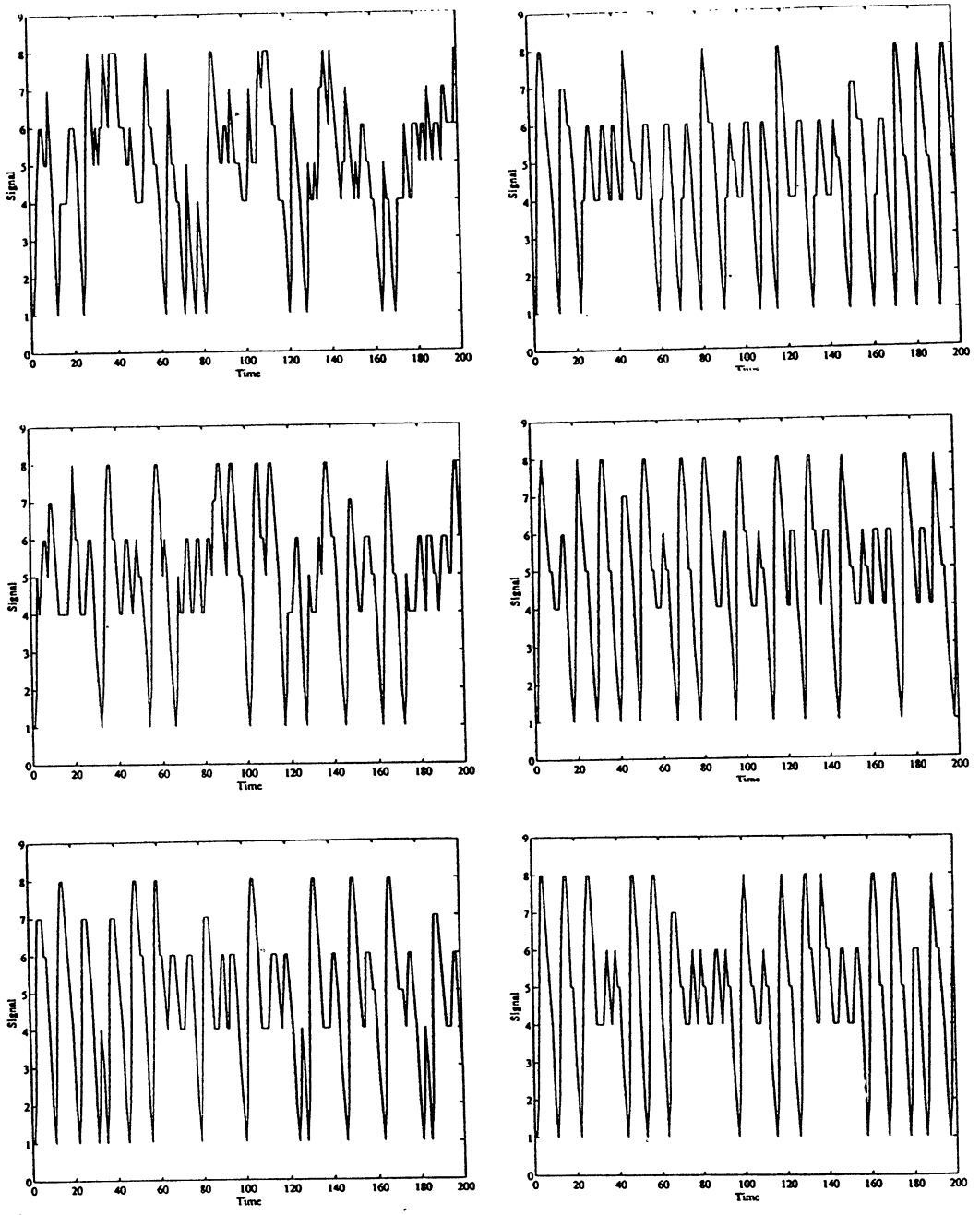
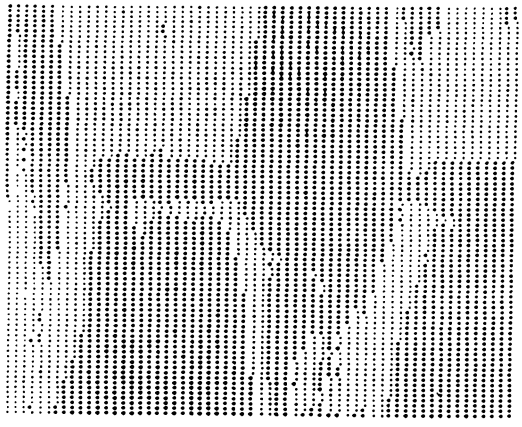
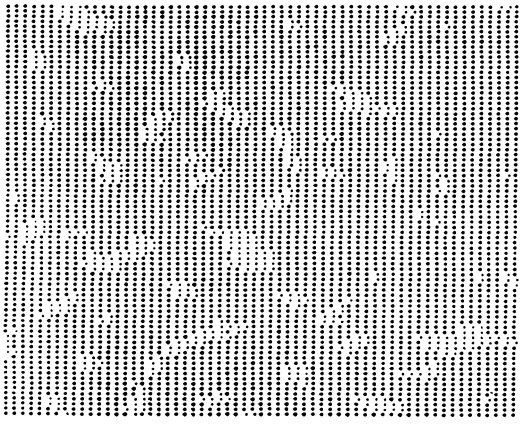


FIG. 3. *HMM's estimated from data in Figure* 1b. *Top to bottom, left-hand side*: 10, 20 *and* 30 *hidden states. Top to bottom, right-hand side*: 40, 50 *and* 60 *hidden states.*



(a)



(b)

FIG. 4. (a) *Thresholded image of straw.* (b) *Thresholded image of paper.*

Following our approach to the speech data, we viewed these images as samples from stationary (spatial) processes, and attempted to fit these processes with N-state hidden Markov models. Specifically, we employed the hiding function $f(x) = x \bmod 2$ and a four-nearest-neighbor Gibbs representation for the hidden process, X_t, $t \in S = \{(i, j)\colon 1 \le i \le 80,\ 1 \le j \le 60\}$. In both experiments, N was fixed at 10, so that $X_t \in \{0, 1, \ldots, 9\}$.

For each texture we fit two matrices $\alpha^h = \{\alpha^h_{kl}\}$ and $\alpha^v = \{\alpha^v_{kl}\}$, where $0 \le k, l \le 9$ and h stands for "horizontal" and v for "vertical." These matrices represent the Gibbs potential for X, as follows:

$$\Pi\left(X_{i,j} = k | X_{i-1,j} = l_1, X_{i+1,j} = l_2, X_{i,j-1} = l_3, X_{i,j+1} = l_4\right)$$
$$\propto \exp - \left\{\alpha^v_{l_1 k} + \alpha^v_{k l_2} + \alpha^h_{l_3 k} + \alpha^h_{k l_4}\right\},$$

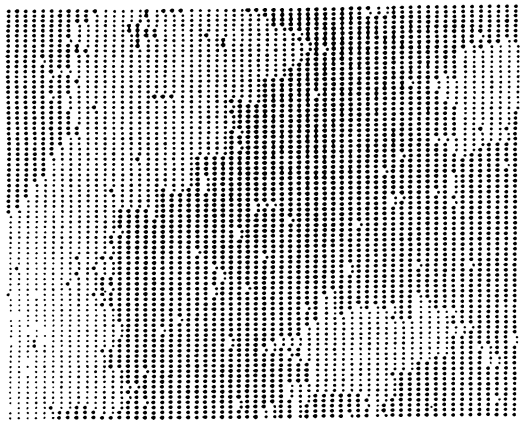
except that terms are dropped when they reference outside of the 80×60 array ("free boundary conditions"). Given a sample $z = \{z_t\}_{t \in S}$, the partial derivative with respect to α_{kl}^h $(0 \leq k, l \leq 9)$ of the log-likelihood of z, under the model $\{f(X_t)\}_{t \in S}$, is

$$E\left[N_{kl}^h\right] - E\left[N_{kl}^h \mid f(X_t) = z_t, t \in S\right], \tag{6}$$

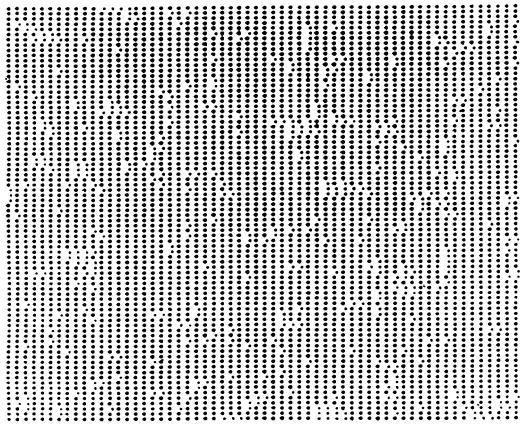
where

$$N_{kl}^h = \#\left\{(i,j)\colon 1 \leq i \leq 80, 1 \leq j \leq 59, X_{i,j} = k, X_{i,j+1} = 1\right\}$$

(a "sufficient statistic"). An analogous expression governs partial derivatives with respect to the components of α^v. One way to estimate the matrices α^h and α^v is via a discrete gradient ascent: compute (6) at the "current" parameter values, take a small step in the direction of the gradient, recom-



(a)



(b)

FIG. 5. (a) *HMM estimated from data in Figure* 4a. (b) *HMM estimated from data in Figure* 4b.

pute (6) and so on. Unfortunately, the computation of (6) is notoriously difficult. We resorted to Monte Carlo methods (cf. Metropolis, Rosenbluth, Rosenbluth, Teller and Teller [39] and Besag and Green [8]), repeatedly using the Gibbs sampler to estimate both expectations.

The approach is unsatisfactory. It is slow and it is difficult to judge convergence, both within an iteration (computation of the expectations) and overall (when to stop?). There have been many suggestions for improving the efficiency of the calculations; see, for example, Younes [45] and Qian and Titterington [42]. We experimented with a variety of alternatives, without much success. In the end we settled on the approach outlined above, which we view as decidedly brute force and last resort.

Having estimated potential functions (α^h and α^v) for both the (binarized) straw and paper textures, we drew samples from the corresponding Gibbs distributions—again, via the Gibbs sampler. The results, viewed through the hiding function f, are shown in Figure 5.

As with the problem of synthesis in speech, texture synthesis is made intriguing by the availability of unlimited amounts of data. Despite this favorable circumstance, there are as of yet no fully satisfactory solutions, especially if one wants to render samples at arbitrary angles and resolution. We have offered a solution, *in principle*: Nearest-neighbor HMM's are dense and can be estimated. Evidently, however, the approach is a long way from being practical. In any case, others have already made good progress: We cite [15], [26], [19], [22], [33] and [21], for some state-of-the-art work on texture estimation and synthesis.

REFERENCES

[1] BAHL, L. R., JELINEK, F. and MERCER, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **5** 179–190.

[2] BALL, F. and SANSOM, M. (1988). Aggregated Markov processes, incorporating time interval omission. *Adv. in Appl. Probab.* **20** 546–572.

[3] BALL, F. G. and RICE, J. A. (1992). Stochastic models for ion channels: introduction and bibliography. *Math. Biosci.* **112** 189–206.

[4] BAUM, L. E. and EAGON, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bull. Amer. Math. Soc.* **73** 360–363.

[5] BAUM, L. E. and PETRIE, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.* **37** 1554–1563.

[6] BAUM, L. E., PETRIE, T., SOULES, G. and WEISS, N. (1970). A maximization technique in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.* **41** 164–171.

[7] BERBEE, H. C. P. and BRADLEY, R. C. (1984). A limitation of Markov representation for stationary processes. *Stochastic Process. Appl.* **18** 33–45.

[8] BESAG, J. and GREEN, P. J. (1993). Spatial statistics and Bayesian computation (with discussion). *J. Roy. Statist. Soc. Ser. B* **55** 25–37.

[9] BICKEL, P. J. and RITOV, Y. (1993). Inference in hidden Markov models I: local asymptotic normality in the stationary case. Technical Report 383, Dept. Statistics, Univ. California, Berkeley.

[10] BILLINGSLEY, P. (1964). *Ergodic Theory and Information*. Wiley, New York.

[11] BLACKWELL, D. and KOOPMANS, L. (1957). On the identifiability problem for functions of finite Markov chains. *Ann. Math. Statist.* **28** 1011–1015.

[12] BRADLEY, R. C. (1993). An addendum to "A limitation of Markov representation for stationary processes." *Stochastic Process. Appl.* **47** 159–166.

[13] BROCKETT, R. W. (1979). Stochastic realization theory and Planck's law for black body radiation. *Ricerche Automat.* **10** 344–362.

[14] BRODATZ, P. (1966). *Texture: A Photographic Album for Artists and Designers.* Dover, New York.

[15] CHELLAPPA, R. and KASHYAP, R. L. (1985). Texture synthesis using 2-D noncausal autoregressive models. *IEEE Trans. Acoust. Speech Signal Process.* **33** 194–203.

[16] CHURCHILL, G. A. (1989). Stochastic models in heterogeneous DNA sequences. *Bull. Math. Biol.* **51** 79–94.

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

[18] DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38.

[19] DERIN, H. and ELLIOTT, H. (1987). Modeling and segmentation of noisy and textured images using Gibbs random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **9** 39–55.

[20] DHARMADHIKARI, S. W. (1963). Sufficient conditions for a stationary process to be a function of a finite Markov chain. *Ann. Math. Statist.* **34** 1033–1041.

[21] ELFADEL, I. M. and PICARD, R. W. (1994). Gibbs random fields, co-occurrences, and texture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **16** 24–37.

[22] FRANCOS, J. M., MEIRI, A. Z. and PORAT, B. (1992). A unified texture model based on a 2-D Wold like decomposition. Technical report, Dept. Electrical Engineering, Technion–Israel Institute of Technology, Haifa, Israel.

[23] FREDKIN, D. R. and RICE, J. A. (1987). Correlation functions of a function of a finite-state Markov process with application to channel kinetics. *Math. Biosci.* **87** 161–172.

[24] Fredkin, F. R. and RICE, J. A. (1992). Maximum likelihood estimation and identification directly from single-channel recordings. *Proc. Roy. Soc. London Ser. B* **249** 125–132.

[25] FRIGESSI, A. and PICCIONI, M. (1990). Parameter estimation for 2 dimensional Ising fields corrupted by noise. *Stochastic Process. Appl.* **34** 297–311.

[26] GAGALOWICZ, A. and MA, S. D. (1985). Sequential synthesis of natural textures. *Computer Vision, Graphics, and Image Processing* **30** 289–315.

[27] GEMAN, S., KEHAGIAS, A. and KÜNSCH, H. (1993). Consistent estimation of stationary processes and stationary random fields. Technical report, Div. Applied Mathematics, Brown Univ.

[28] GEORGII, H.-O. (1988). *Gibbs Measures and Phase Transitions.* de Gruyter, New York.

[29] GILBERT, E. J. (1959). On the identifiability problem for functions of finite Markov chains. *Ann. Math. Statist.* **30** 688–697.

[30] GRENANDER, U. (1978). *Abstract Inference.* Wiley, New York.

[31] ITÔ, H., AMARI, S.-I. and KOBAYASHI, K. (1992). Identifiability of hidden Markov information sources and their minimum degrees of freedom. *IEEE Trans. Inform. Theory* **38** 324–333.

[32] JI, C. (1990). Sieve estimators for pair-interaction potentials and local characteristics in Gibbs random fields. Technical Report 2037, Institute of Statistics, Univ. North Carolina.

[33] JI, C. and SEYMOUR, L. (1992). On the selection of Markov random field texture models. Technical report, Dept. Statistics, Univ. North Carolina.

[34] KALLIANPUR, G. (1980). *Stochastic Filtering Theory.* Springer, New York.

[35] KEHAGIAS, A. (1992). Approximation of stochastic processes by hidden Markov models. Ph.D. dissertation, Div. Applied Mathematics, Brown Univ.

[36] KIENKER, P. (1989). Equivalence of aggregated Markov models of ion-channel gating. *Proc. Roy. Soc. London Ser. B* **236** 269–309.

[37] KINDERMANN, R. and SNELL, J. L. (1980). *Markov Random Fields and Their Applications*. Amer. Math. Soc., Providence, RI.

[38] KROGH, A., BROWN, M., MIAN, I. S., SJÖLANDER, K. and HAUSSLER, D. (1994). Hidden Markov models in computational biology: applications to protein modeling. *Journal of Molecular Biology* **235** 1501–1531.

[39] METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21** 1087–1092.

[40] ORNSTEIN, D. S. and WEISS, B. (1990). How sampling reveals a process. *Ann. Probab.* **18** 905–930.

[41] PINCUS, S. M. (1992). Approximating Markov chains. *Proc. Nat. Acad. Sci. U.S.A.* **89** 4432–4436.

[42] QIAN, W. and TITTERINGTON, D. M. (1991). Estimation of parameters in hidden Markov models. *Philos. Trans. Roy. Soc. London Ser. A* **337** 407–428.

[43] RABINER, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77** 257–286.

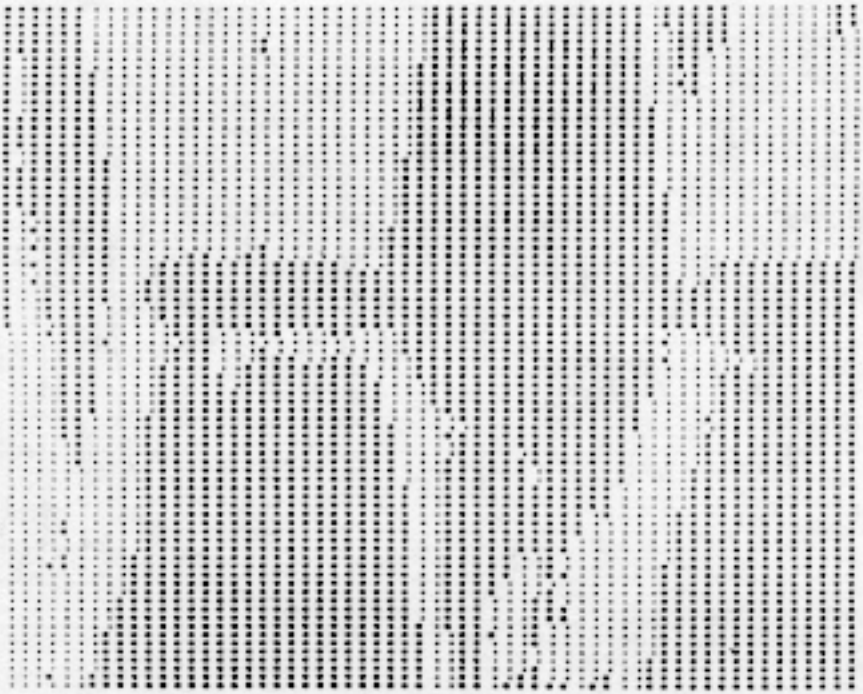
[44] ROSENBLATT, M. (1971). *Markov Processes. Structure and Asymptotic Behavior*. Springer, New York.

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

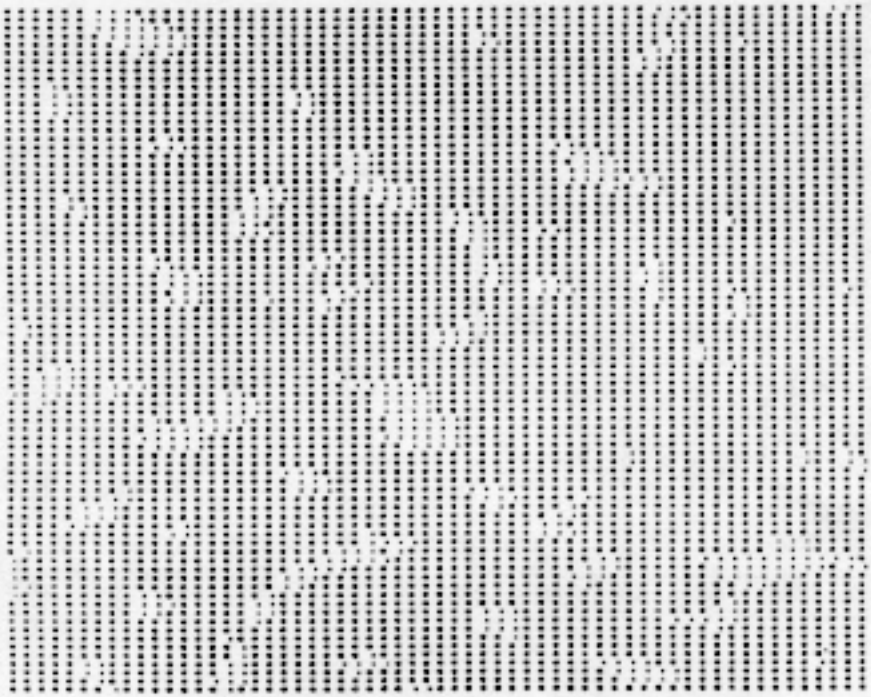
HANS KÜNSCH
SEMINAR FÜR STATISTIK
ETH ZENTRUM
CH-8092 ZURICH
SWITZERLAND

STUART GEMAN
DIVISION OF APPLIED MATHEMATICS
BROWN UNIVERSITY
PROVIDENCE, RHODE ISLAND 02912

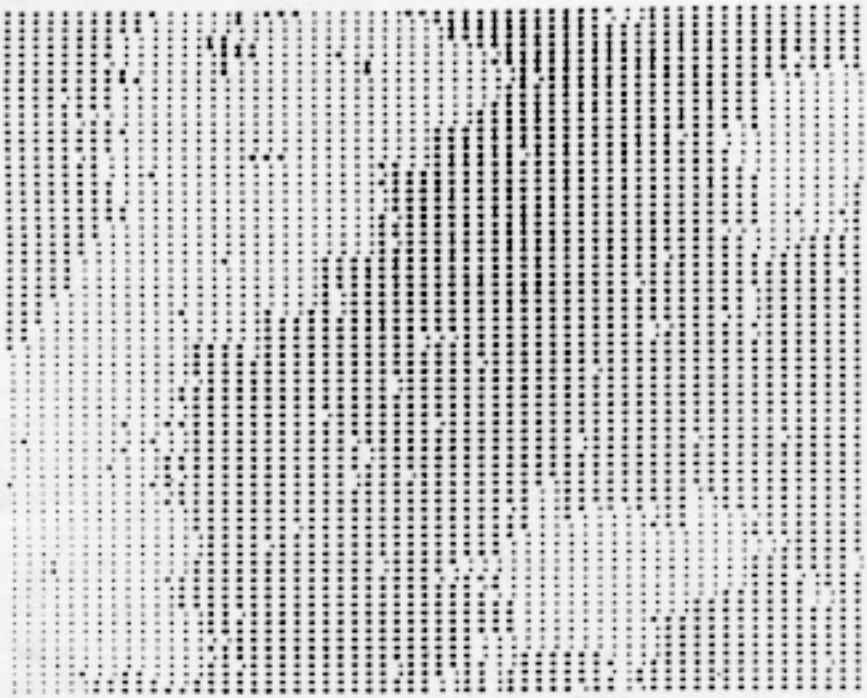
ATHANASIOS KEHAGIAS
DIVISION OF ELECTRONICS AND COMPUTERS
ELECTRICAL ENGINEERING DEPARTMENT
ARISTOTLE UNIVERSITY OF THESSALONIKI
GREECE



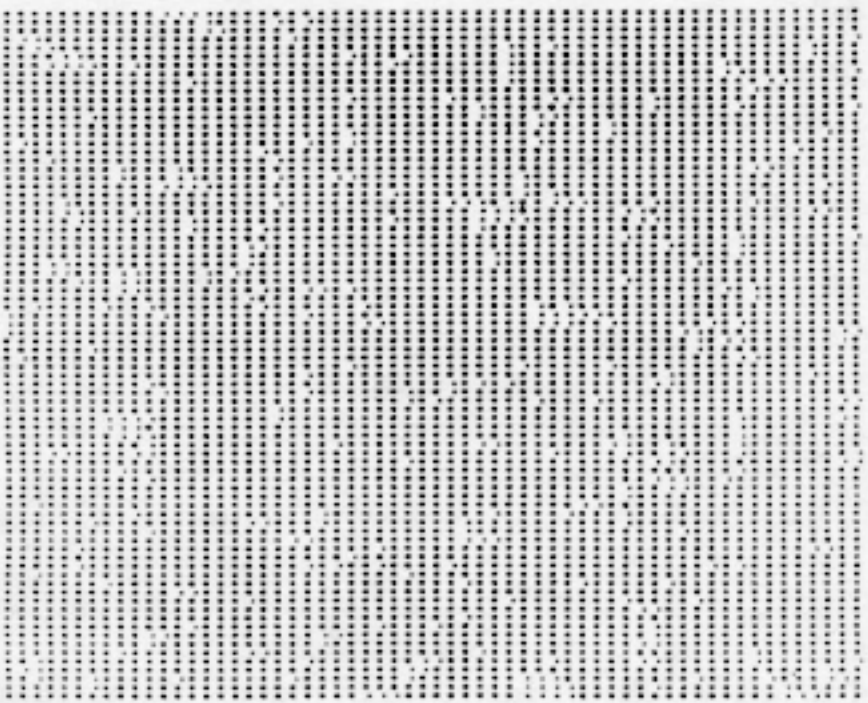
(a)



(b)



(a)



(b)