The Annals of Statistics
1989, Vol. 17, No. 1, 1–30

THE 1985 RIETZ LECTURE

ADVANCES IN PATTERN THEORY[1]

BY ULF GRENANDER

Brown University

Pattern theory offers concepts for modelling images and methods for making inferences from observed images. This will be described briefly and illustrated by examples.

We shall present limit theorems for the Markov processes on graphs (that are basic to pattern theory) motivated by computational considerations. They will yield approximations that have been exploited to make the inference algorithms computationally feasible.

We shall also consider the problem of estimating parameters in the prior measures encountered in pattern theory. These parameters are high dimensional, not automatically identifiable and notoriously difficult to estimate by standard methods. We therefore present a standardization technique for dealing with them and show how, after standardization, the remaining free parameters can be estimated by different methods. The estimation methods are examined in terms of their asymptotic efficiencies.

1. Inference machines for parallel logic. We shall study inference machines that implement *parallel logic under uncertainty for complex systems*. The underlying sample space describing such systems will be defined and analyzed in terms of general pattern theory and we shall summarize the fundamental concepts and methods and illustrate them by examples. The interested reader will find more detailed information about this in Grenander [(1976, 1981a), Volume I, Sections 1.1, 2.1–2.2, 3.1 and Volume III, Sections 3.1–3.2 and Chapter 5].

1.1. The sample space C—the *configuration space*—will be made up of mathematical objects generically denoted by $c = \sigma(g_1, g_2, \dots, g_n)$—the *configurations*.

Here the g_i, $i = 1, 2, \dots, n$, are elements from a *generator space* G. To each $g \in G$ is associated a number $\omega(g)$, the *arity* of g. To each $j = 1, 2, \dots, \omega(g)$ is attached a *bond value* $\beta_j(g)$ with elements from a *bond value space* B.

The *connector* graph σ over $|\sigma| = n$ sites expresses how the generators g_i communicate with each other and will represent segments between sites i_1 and i_2 of the connector as in Figure 1. The generator g_{i_1} sends a signal $\beta_1 = \beta_{j_1}(g_{i_1})$ that meets the signal $\beta_2 = \beta_{j_2}(g_{i_2})$ emanating from site i_2. Depending upon how β_1 and β_2 "agree," in a sense to be defined later, g_{i_1} and g_{i_2} will influence each other directly in deterministic or probabilistic terms.

Received February 1987; revised September 1988.

[1] Supported by U.S. Army Research Office Contract DAAG-29-83-K-0116 and National Science Foundation Grant MCS-83-06507.

AMS 1980 *subject classifications*. Primary 60B99; secondary 62M02.

Key words and phrases. Pattern theory, image processing, limit theorem, estimation of acceptor functions.

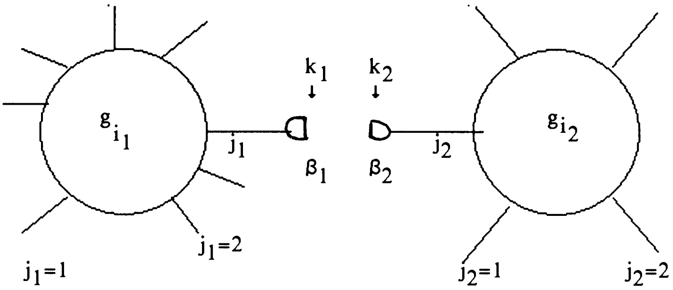

FIG. 1. *Interacting generators.*

Only generators at sites connected by a segment in σ are influencing each other directly. Other pairs may also exert influence on each other but only indirectly via paths in the graph. In the figure the bond $k_1 = (i_1, j_1)$ connects with the bond $k_2 = (i_2, j_2)$ forming a segment $s = (i_1, i_2) \in \sigma$.

In general the connector σ need not be fixed; instead we specify a set Σ, the *connector type* of graphs and ask that $\sigma \in \Sigma$. For example, Σ may be the set of all rooted trees as in Section 1.2.1.

We shall only assume pairwise interaction in the following. Actually this can be shown to be no essential restriction, but we shall not discuss this topic here.

1.2. Given a truth valued function ρ, the *bond relation* ρ: $B \times B \to \{\text{true}, \text{false}\}$ we shall denote by $C(\mathscr{R}) \subseteq C$ (where $\mathscr{R}$ is defined below) the set of all configurations in C such that the bond relation is satisfied along all segments of σ. $\rho[\beta_{j_1}(g_{i_1}), \beta_{j_2}(g_{i_2})] = \text{true}$ for all segments $s = (i_1, i_2) \in \sigma$. We can express this by the *structure formula*

$$(1.1) \qquad \bigwedge_{(i_1, i_2) \in \sigma} \rho\left[\beta_{j_1}(g_{i_1}), \beta_{j_2}(g_{i_2})\right] = \text{true}$$

for the regularity $\mathscr{R} = \langle G, \sigma, \rho \rangle$. Here the graph σ represents the *logical architecture* and the *global regularity*, the the bond relation ρ expresses the *local regularity*.

On the sample space C we introduce a *prior measure* P by another structure formula,

$$(1.2) \qquad P(c) = \frac{1}{Z} \prod_{(i_1, i_2) \in \sigma} A\left[\beta_{j_1}(g_{i_1}), \beta_{j_2}(g_{i_2})\right] \prod_{i=1}^{n} Q(g_i)$$

for finite generator space $|G| = r < \infty$. If G is infinite, for example $\mathbb{R}^d$, we interpret the right-hand side of (1.2) as a density with respect to some fixed measure.

In (1.2) A is the *acceptor function* A: $B \times B \to \mathbb{R}^+$, Q is a weight function Q: $G \to \mathbb{R}^+$ and Z is a normalizing constant, the *partition function* well known in statistical mechanics, ensuring that $P(C) = 1$. Z is notoriously difficult to calculate except for the simplest cases.

Probability models related to (1.2) have appeared in many contexts, of which we mention genetics, physics and pattern theory.

In the case of *rigid regularity*, which means that $\rho(\beta_1, \beta_2) = 0$ (meaning false) implies $A(\beta_1, \beta_2) = 0$, the support of P is in $C(\mathscr{R})$. If this is not the case we speak of *relaxed regularity*. In either case we write the regularity $\mathscr{R} = \langle G, \sigma, A \rangle$.

Since the measures in (1.2) are induced by the regularity $\mathscr{R}$ one speaks of *regularity controlled probabilities*.

1.2.1. To make the above more intuitive we shall start by an example from formal language theory: *context free grammars*, structures that have been studied for a long time.

Consider a finite vocabulary

$$V = V_T \cup V_N,$$

where $V_T = \{a, b, c, \dots\}$ is the terminal vocabulary containing "words" and $V_N = \{\alpha, \beta, \dots\}$ is the nonterminal vocabulary containing syntactic variables (such as noun, verb phrase, article, ...). In V_N we let α be the initial variable that will be at the apex of all derivation trees.

We also need a finite set W of rewriting rules. Each is of the form

$$r_i \rightarrow s_{i1}, s_2, \dots, s_{ij} \in V, \qquad r_i \in V_N.$$

We start with α and rewrite each nonterminal appearing in the derived string using rewriting rules repeatedly until the derived string contains only terminals. Then we stop.

A derivation could look like Figure 2a. We have used the abbreviations NP = noun phrase, VP = verb phrase, ART = article, ADJ = adjective, N = noun and V = verb. The derived sentence is "the big dog saw the little cat." The set of all finite strings that can be derived, the syntactically correct sentences, is called the language generated by the system.

This is naturally identified with a regular structure as follows:

G: generators are the rewriting rules, each of which has one in bond (a syntactic variable) and a finite number of outbonds (from V) so that the bond

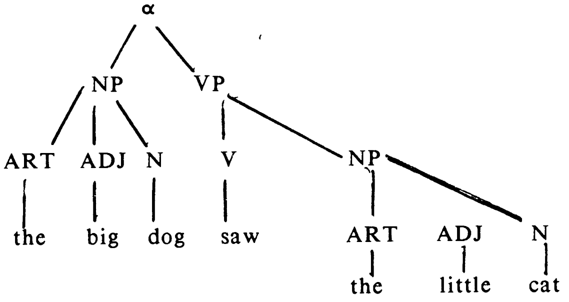

FIG. 2a. *Derivation of sentence.*

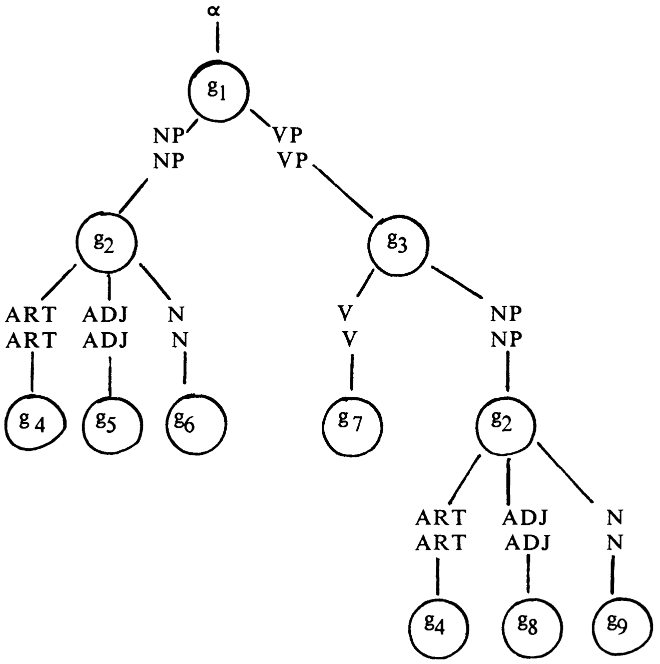

FIG. 2b. *Configuration diagram.*

value space

B: equals the vocabulary V.

Σ: the set of connector graphs consists of rooted trees.

ρ: the bond relation means equality.

C: the configuration space consists of all syntactically correct phrases (subtrees of correct derivation trees).

A more complete configuration diagram is given in Figure 2b with the rewriting rules

g_1: $\alpha \to$ NP, VP

g_2: NP → ART, ADJ, N

g_3: VP → V',NP

g_4: Art → the

g_5: ADJ → big

g_6: N → dog

g_7: V → saw

g_8: ADJ → little

g_9: N → cat

1.2.2. Our next example appears in the global shape models that have been used to describe three-dimensional objects. Let us start with the biggest of the five regular polyhedra: the icosahedron. Each of its faces is an equilateral triangle. Subdivide each face into four naturally congruent equilateral triangles. Repeat this procedure l time so that we have a finely tesselated polynedron. A homeomorphic map of it could look like the one in the middle of Figure 3 (due to D. Keenan); the four others represent samples from a regularity controlled prior measure.

G: the edges of the polyhedron.

B: the bonds of a generator should be the endpoints of the edge, so that $B = \mathbb{R}^3$.

Σ: the set of connector graphs will be the edge graph of the polyhedron (where two edges are said to be neighbors if they share one endpoint) and its subgraphs.

ρ: here the bond relation means equality.

C: the set of subpolyhedra of the lth level polyhedron.

1.2.3. As a third example let us consider schemas made up of computational modules. Each computing module represents a computing operation, a function. Its input $x = (x_1, x_2, \dots, x_k)$ has components $x_1, x_2, \dots$ with values in spaces $X_1, X_2, \dots$, respectively. Its output $y = (y_1, y_2, \dots, y_l)$ has components $y_1, y_2, \dots$ with values in spaces $Y_1, Y_2, \dots$. The values of k or l may be zero.

Combining the modules by a wiring diagram we must make sure that a function is not called for execution before its inputs have been evaluated. We do this by arranging the modules by a partial order, assuming that all looping, if any, takes place within modules.

We must also make sure that outputs belong to input spaces for two connected modules. An example is given in Figure 4 that illustrates schematically a computation schema for the function

$$y_1 = 4(\log x_1 + 1/x_2),$$

$$y_2 = \sqrt{x_1 + x_2^2},$$

where we assume all variables to be real, $x_1 > 0$, $x_2 \neq 0$.

The pattern theoretic formalism for this can then be chosen as:

G: set of functions (computing modules).

B: the bonds are the X and y spaces of the inputs/outputs of a function.

Σ: Poset graphs are the connectors.

ρ: the bond relation means inclusion.

C: the configuration space represents computing schemata.

In this case the prior could be given, in the notation of (1.2), by $A(\beta', \beta'') = \delta_{\beta'\beta''}$ and some weight vector Q with values < 1. This would make large configurations less likely than smaller ones.

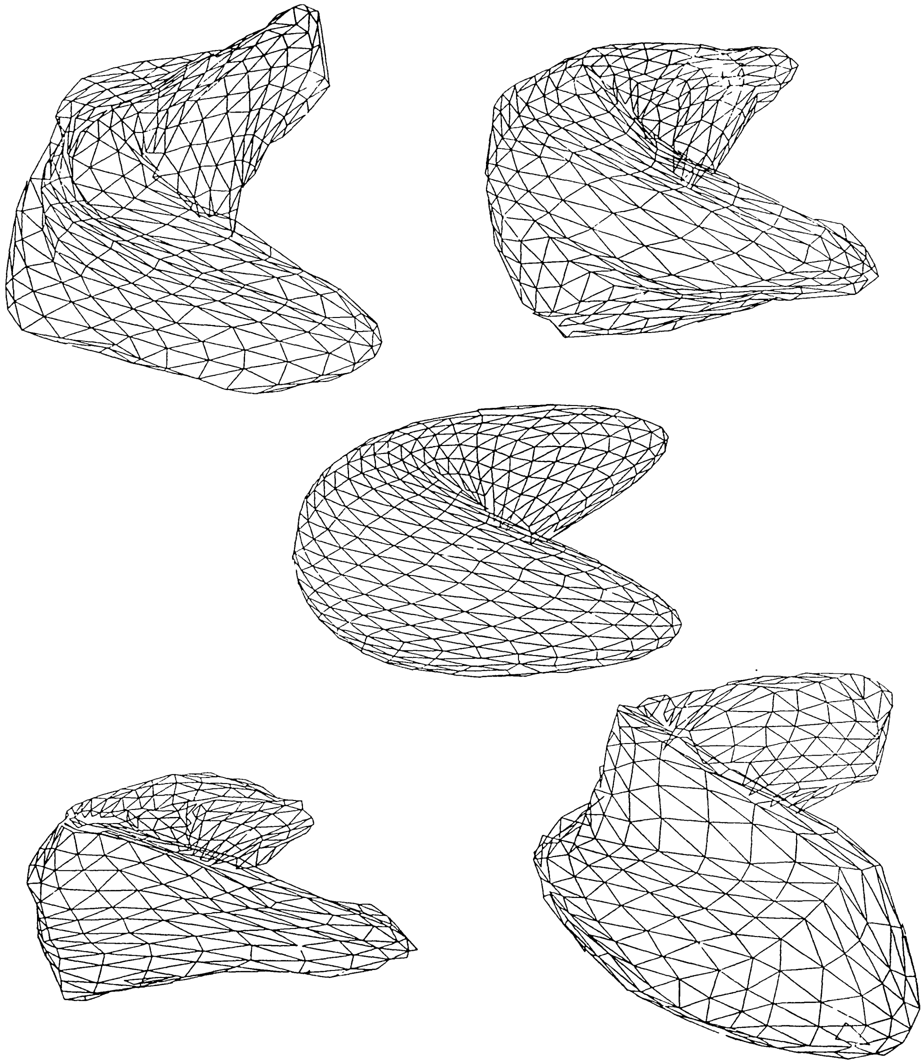

FIG. 3. *Random shapes in* $\mathbb{R}^3$.

1.3. It is well known that P defines a *Markov process in the graph* σ in the following sense. Consider Figure 5 as an example. We then have, assuming for simplicity that A and Q are strictly positive,

$$P(c'|c'') = P(c'|c'''), \tag{1.3}$$

so that the conditional probability of a subconfiguration c' given the rest c'' of c equals the conditional probability of c' given the *boundary* c'''.

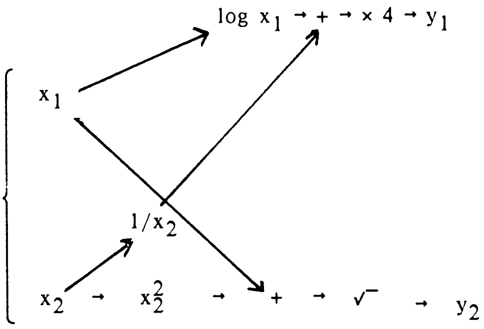

FIG. 4. *Computational configuration.*

A similar but stronger statement is due to Thrift (1979):

THEOREM 1. *Given a subconfiguration $c' \subset c$ and a site $i \in c$, we have*

$$P(g_i|c') = P(g_i|c'') \tag{1.4}$$

where c'' consists of all sites that are connected to i by a chain outside (i.e., all its sites are in the complement of c') c'.

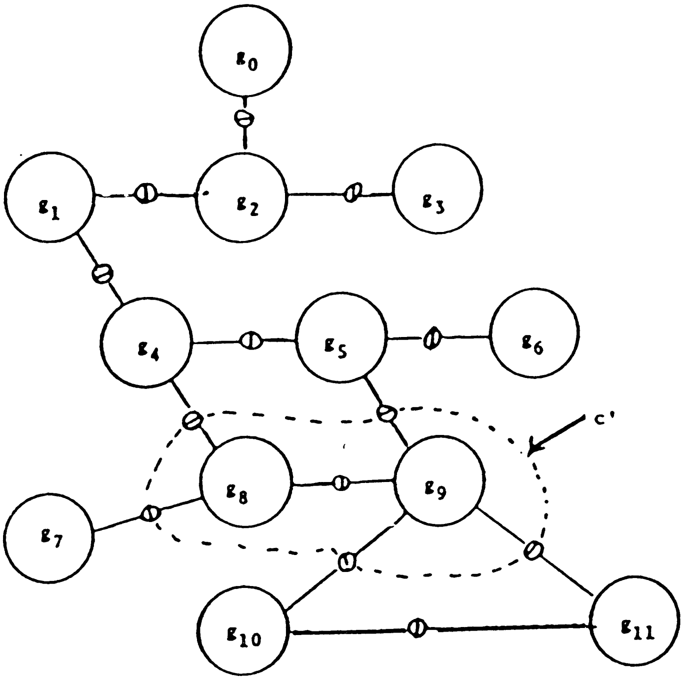

FIG. 5. *Configuration with subconfigurations.*

A proof can be found in Grenander [(1981), Section 5.2].

We shall write

$$A(B_1, B_2) = \exp[a(\beta_1, \beta_2)],$$

$$(1.5) \qquad a(B_1, B_2) = \frac{1}{T} a_0(\beta_1, \beta_2) \quad \text{or, if } B \text{ is a vector space,}$$

$$a(\beta_1, \beta_2) = a_0\left(\frac{\beta_1 - \beta_2}{\varepsilon}\right)$$

in terms of the *affinities* $a(\beta_1, \beta_2)$ and the *temperature* T. The temperature measures the degree of disorder in the relaxed regularity. The role of the scale parameter ε is similar to that of T.

In general the acceptors A may be allowed to depend upon the segment $s = (i_1, i_2)$. We then indicate the dependence by the notation A_s.

1.4. The study of these complex systems is made harder by the fact that the configurations are often *not completely observable*. Instead the observer can only see some function $I = R(c)$ of the configuration c, where R, the *identification rule*, is not generally invertible. The possible images form a set $\mathscr{T}$, the so-called *image algebra*; see Grenander [(1976), Chapter 2] where they are studied as partial universal algebras. The prior measures in (1.2) can then be viewed in the context of probabilities on algebraic structures, but this will not be pursued in the present paper.

In some, but not all cases to be discussed below, R is just a function applied to each generator g_i separately so that the *image* I consists of observations $R(g_1), R(g_2), \dots, R(g_n)$. It is clear that we cannot in general claim that I forms a Markov process on σ. Instead our prior measures will be *incompletely observed Markov processes on graphs.*

Let us see what identification rule is natural in the above three examples. In 1.2.1, two configurations will be identified if the resulting strings are equal. Note that for some context free grammars a string may be derived in more than one way, so that R will be many-to-one. The image I represents a syntactically correct phrase, not just its derivation(s).

In Section 1.2.2, we shall identify two polyhedra if they represent the same geometric object in $\mathbb{R}^3$. We then lose knowledge of the subscripting of the edges and R is again many-to-one.

In Section 1.2.3, we shall identify two configurations if they represent the same function. A configuration is a meaningful formula, and many formulas may represent one function.

1.5. A further complication is that the image I may also be *hidden by noise*. A stochastic deformation mechanism $\mathscr{D}$ operates on I, resulting in some deformed image $I^{\mathscr{D}} = \mathscr{D}I$. Perhaps $\mathscr{D}$ consists of additive noise, symmetric binary noise, blurring (convolutions) or a mask operation hiding part of the image.

The typical inference problem then looks like the following. Having observed a deformed image $I^{\mathscr{D}}$, reconstruct the pure image I by a procedure that is optimal in some given sense. Since we have a prior measure on $\mathscr{T}$ we use a Bayesian approach.

The dimensionality of the image algebra $\mathscr{T}$, which here plays the role of parameter space, can be enormous, easily 10^3, even 10^6 in some image processing applications. We are therefore dealing with problems in *abstract inference*. In addition to the Bayesian approach described in this paper one can also use the method of sieves [see Grenander (1981b), Part III] but this possibility will not be studied here.

1.6. For Bayesian inference we need the posterior measure

$$P(I|I^{\mathscr{D}}) = \frac{P(I)P(I^{\mathscr{D}}|I)}{P(I^{\mathscr{D}})}. \tag{1.6}$$

Unfortunately (1.6) is usually awkward to handle analytically due to the appearance of the partition function Z in (1.2). Instead we shall solve the inference problem by *simulating* (1.6) *for a fixed observed* $I^{\mathscr{D}}$ and produce a sample from this distribution. For a given optimality criterion we then use the sample to construct an estimate I^*.

For high regularity so that $A(x, x')$ is small unless x is close to x', it may be enough to use sample size 1 since the posterior in (1.6) is then typically very peaked.

But how can one simulate Markov processes on graphs? Direct simulation seems possible only if σ has no cycles—it is a tree. Then one can, at least in principle if not always in practice, start simulating g_{i_1} for some site i_1 and then follow the graph from i_1 simulating conditional distributions until all the sites in σ have been dealt with.

The case when σ has cycles is the situation when parallel logic is of greatest interest: *Conflicting evidence will have to be resolved.* We can then use a modification of the *Metropolis algorithm*; see Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953). The basic idea is to use the Markov property for each site and it goes as follows.

Step 1. Initialize c according to some convention.
Step 2. Initialize by setting $i = 1$.
Step 3. At site i simulate g_i conditioned by the neighboring sites.
Step 4. Update i to $i + 1$, $i < n$ or 1 if $i = n$.
Step 5. Go to Step 3 or stop when the number of iterations is deemed sufficient.

It is easily seen that the Markov chain $C_1, C_1, C_3, \ldots$ with the state space C, assuming that A and Q were assumed to be strictly positive, is ergodic and that the above procedure, *stochastic relaxation*, converges to the unique equilibrium measure that coincides with (1.6).

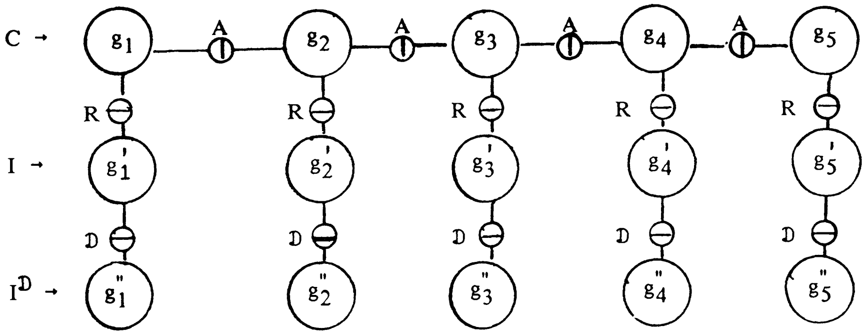

FIG. 6. *Inference machine.*

The sequence $1, 2, \ldots, n, 1, 2, \ldots$ can be replaced by other *sweep strategies* as long as they guarantee that each site will be updated i.o.

1.7. To visualize stochastic relaxation consider Figure 6 in which for simplicity σ is just a linear graph with $|c| = 5$.

In this three-level *inference machine* the upper level represents the configuration $c = \sigma(g_1, g_2, \ldots, g_n)$ following the regularity $\mathscr{R} = \langle G, \sigma, A \rangle$, the second level the (pure) image $I = Rc$ and the third level the observed image $I^{\mathscr{D}} = \mathscr{D}I$.

During the stochastic relaxation all the third level generators g_i'' are kept fixed to the values observed, while the others are updated by visiting them following some sweep strategy.

The resulting estimate I^*, the image restoration, appears in the second level after the relaxation algorithm has been executed. The configuration c^* in the first level is the synthesis, or explanation, of I^* and expresses our *understanding* of I^*.

1.8.1. Let us make the above more concrete by the following three examples. In the first we choose $G = \{0, 1\}$, σ a cyclic square $L \times L$ lattice, $n = L^2$, R is the identity and $\mathscr{D}$ represents a noisy binary, symmetric channel with error rate ε. This is just an instance of the celebrated *Ising model* of ferromagnetism; see, e.g., Kinderman and Snell (1980). As a model for picture processing it lacks enough generative power and is only used here because of its simplicity.

In Figure 7 we show the pure image I (here coinciding with the configuration c), the deformed $I^{\mathscr{D}}$ with $\varepsilon = 20\%$ and the restored image I^*; the latter was obtained by the procedure described in Section 1.6.

1.8.2. In our second example σ is still a cyclic square lattice but now with arity eight for the sites so that each site has eight neighbors. The generator space now is of size $r = |G| = 42$ (if redundancy is removed deleting identical generators), with more structured elements representing *geometric tendencies*.

For example $g = 0$ expresses the tendency for a site to become an outside point, $g = 1$ an inside point and the remaining 40 generators represent tenden-

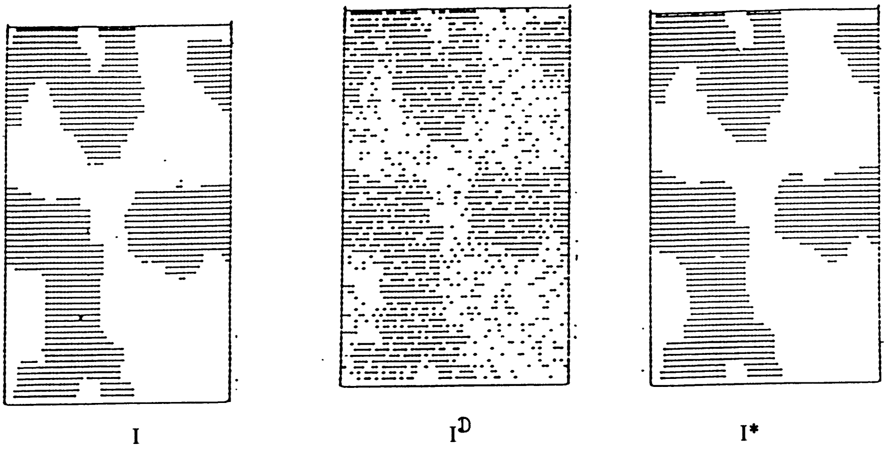

FIG. 7. *Restoration of binary image.*

cies to become boundary elements with different curvatures and orientations. This information is coded into the eight bond values for each g. Space does not allow us to describe the details that can be found in Grenander [(1983), pages 109–115].

Here R maps $g = 0$ into 0 (white) and all other g's into 1 (black), while $\mathscr{D}$ is the same as in the first example.

In Figure 8 we show a pure image in (a), a deformed image in (b) and the restored one in (c), the latter obtained by stochastic relaxation of the posterior.

1.8.3. In the third example where the images will be random polygons the generators consist of line segments and ρ expresses the condition that the sides of the resulting polygon should connect at endpoints. The function Q attaches weights to segments of different lengths and directions. Note that the connector graph σ now consists of one or several cyclic chains; we shall consider a single one here.

So far this describes the *shape*, and it remains to decide on the location, orientation and size of the polygon. We do this by specifying a probability measure over the group consisting of the Euclidean group in R^2 times the uniform scale change group. In the computer experiment we have simply used simple unform measures.

Note that in this example the prior is made up of two measures: a measure of the form (1.2) for shape and another measure to account for location, orientation and scale.

The identification rule R interprets such a configuration as the inside of the corresponding polygon.

In Figure 9 we show the result of a computer experiment using the regular structure just described.

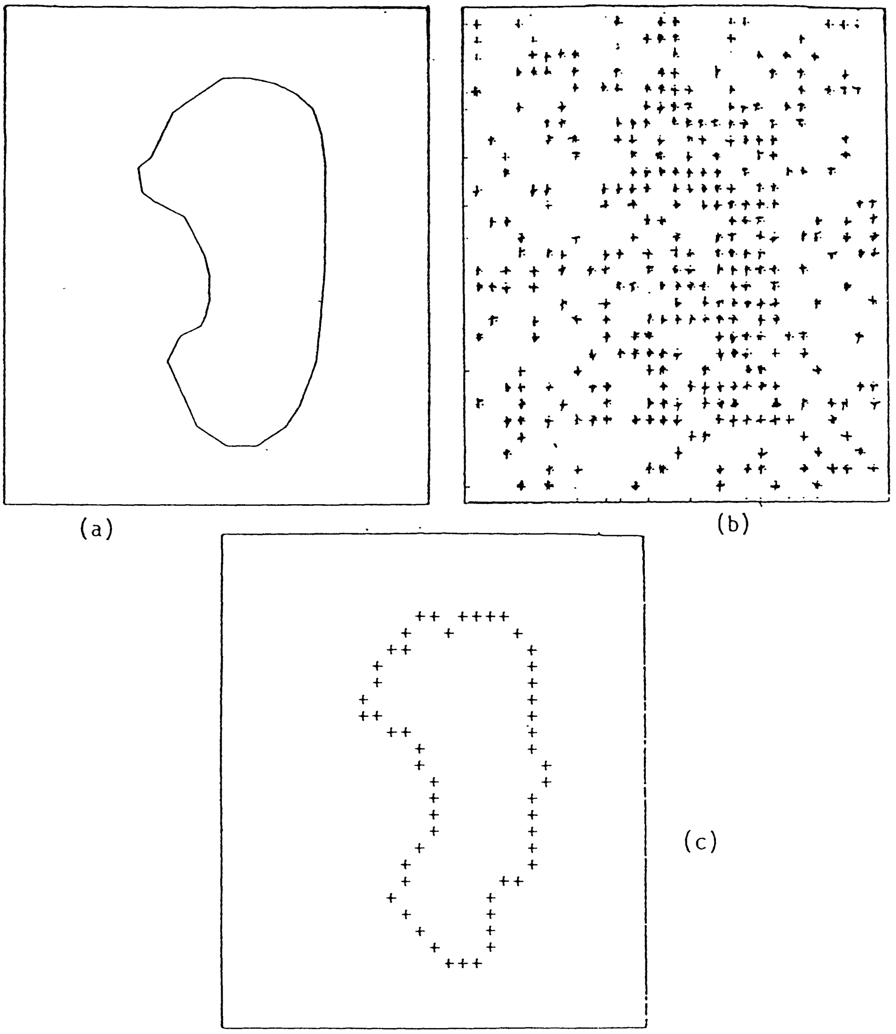

FIG. 8. *Restoration with boundary generators.*

Let us mention in passing that this is a special case of an emerging *theory of random shape*; see Grenander and Keenan (1988). It deals with objects in $\mathbb{R}^2$, and, in a tentative manner with objects in $\mathbb{R}^3$, and will be presented elsewhere.

The estimation procedure obviously depends on what optimality criterion is used. One attractive possibility is to search for the I that maximizes (1.6), the mode in the posterior.

Since the posterior can have many local maxima, standard hill climbing techniques are not recommended. Instead *simulated annealing* seems reasonable, lowering the temperature [see (1.5)] slowly during the stochastic relaxation. The question of the *rate of decreasing temperature* was recently answered; see

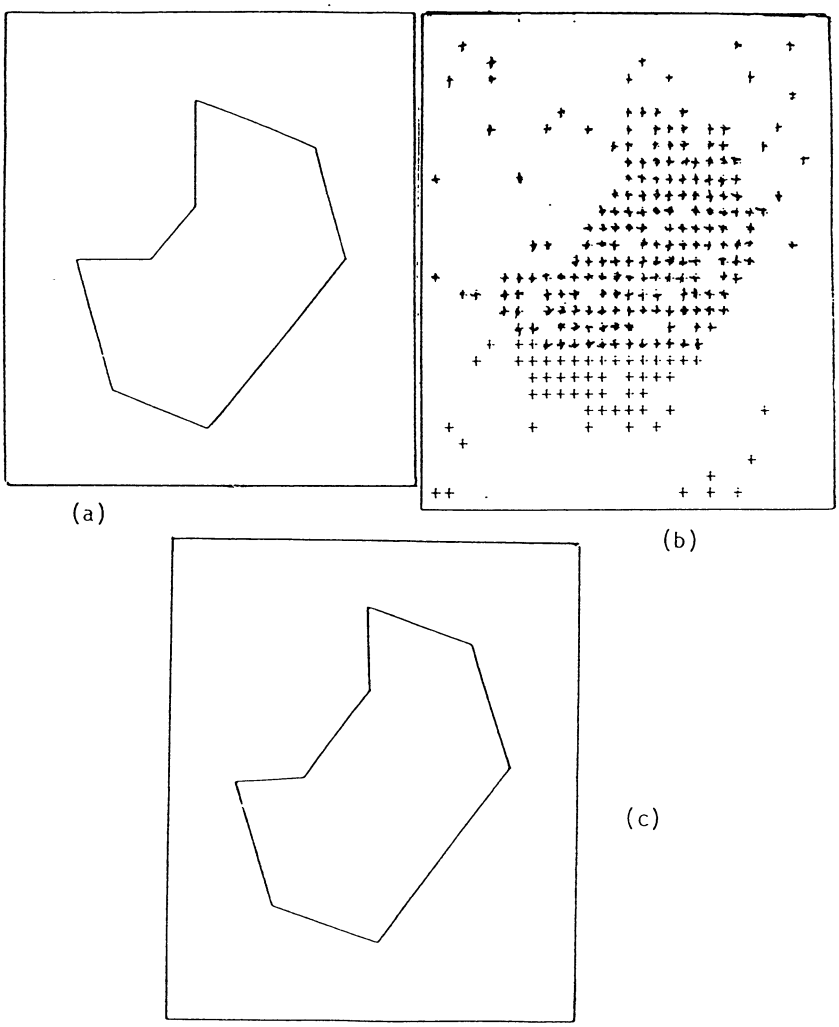

FIG. 9. *Restoration with global shape model.*

Geman and Geman (1983) and Gidas (1985). Convergence to the mode is guaranteed if the temperature T is lowered at a rate not faster than constant/log t, where t is the iteration number. This has been used with success in a large number of image processing computer experiments.

It is not known what the best procedure is in practice. Based only on computational experience, but without analytical backup, the author prefers to simulate the posterior a number of times and "average" (in some sense, not necessarily linear) the results. This corresponds to estimating some posterior "mean" rather than the mode. The latter may not always be representative of the posterior in the sense of being close to where most of the probability mass is.

1.8.4. These and many other similar experiments indicate that excellent restoration can be achieved. But this is not the main point. One can get good restoration by more ad hoc methods. However, it is not possible to discuss in precise terms how good a restoration is, how well it compares with a theoretically optimal method, unless one starts from a *model based* approach.

Here we are doing just that, building the models of random geometries on pattern theoretic ideas and then *deriving* the optimal algorithm (for a given optimality criterion).

This is clearly preferable *if* it can be done, and it should be pointed out that serious obstacles must be removed in order to carry this out. Let us mention some of these difficulties.

1.8.4.1. First, and most importantly, we need a *repertoire of pattern theoretic* models (in the case of picture processing random geometries, in particular random shape models) from which we can select a suitable model. Random geometries have been studied for a long time in integral geometry [see also Harding and Kendall (1974)], but we need more specific models. This is not primarily an analytical problem: It is not a well-posed mathematical problem; it requires more intuition and inventiveness than analytical skills. It will not be discussed here. The interested reader can find a large number of pattern theoretic models in Grenander (1976, 1978, 1981a) and Grenander (1983) but much more is needed.

1.8.4.2. Once the structure of the model has been determined we must decide the values for its parameters. It is only recently that (partial) answers have been obtained to the question of parameter estimation for these models. This will be discussed in Section 2.

1.8.4.3. After completely specifying the model we implement it, for example, by stochastic relaxation as described. It has become clear that the *relaxation time can be large*, so that massive computation is needed

(a) if the graph is large, n big,
(b) if the regularity is high, T or ε small,
(c) if $|G|$ is big or even infinite,
(d) if the conditional measures are not of familiar form.

In particular we must expect extreme CPU time requirement if all of (a)–(d) hold.

Recent advances in *computer architecture* tend to ameliorate the situation; in particular *array and parallel architecture* seem tailormade for implementing the mathematics of parallel logic/complex systems. Nevertheless, there is a need for analytical improvements, in particular for limit theorems that yield more tractable approximations to the probability measures. This will be treated in Section 3.

1.9. Before presenting some new analytical results let us briefly mention some possible applications. A massive research effort is under way with the goal of applying the ideas to image processing. Other applications are being discussed, so far only in a *speculative* manner.

1.9.1. *Computer aided medical diagnosis*, usually discussed from the perspective of expert systems and knowledge engineering, seems a natural application. The sites would carry information about patient data, personal characteristics, test results, medical history, disease × on/off, etc. The connector would express the current state of medical doctrine, one site is believed to influence another site directly but a third site only indirectly and so on. The strength of interactions would be expressed by the acceptor functions.

Some sites would be observed, others would be estimated by the inference machine, and determined by conditional probabilities.

1.9.2. *Decision making in complex systems* where there is a lot of data, vague and uncertain, some of it hidden. Again, the goal would be to model the complex system in pattern theoretic terms by a regular structure and implement it by an inference machine. Of course the acceptor functions must be determined empirically.

1.9.3. Biometric situations, say related to *pharmacology*, in which a large number of treatments (substances) are tested under various conditions. Model building in such a priori unstructured situations can be expected to be difficult, but Markov processes on graphs, perhaps incompletely observed, form a tool worth trying. One obstacle, appearing in the absence of an explicit subject matter theory, is the choice of the graph σ.

1.9.4. Analysis of biological shapes, for example, the shape of hands and other biological forms [Grenander and Keenan (1988)] and their automatic recognition by computer. This is currently being done.

2. Estimating acceptors. When it comes to estimating unknown parameters in the acceptor functions two cases should be distinguished. In the first all acceptor functions are the same, there is only a single one, and one treats large graphs. The asymptotics then means that the size of the graph, n, tends to infinity. In the second case we allow acceptor functions to vary from edge to edge. The natural asymptotics then is to assume that the number of observed images, N, tends to infinity.

Let us try to determine the unknown entries in the acceptor matrices A_s empirically. Say that $|G| = r < \infty$ and that we have observed an i.i.d. sample of configurations $c(1), c(2), \dots, c(t), \dots, c(N)$ from P given by (1.2). For simplicity we shall assume "full information bonds," i.e., $\beta_j(g) = g$. First it is clear that we can absorb the Q-factor into the A_s's without restricting the family of prior probability measures. We shall write, for the time being,

$$P[c(t)] = \frac{1}{Z}\exp\left[\sum_{s=1}^{m} a_s(g_{i_1}, g_{i_2})\right], \qquad t = 1, 2, \dots, N, \tag{2.1}$$

and

$$c(t) = \sigma[g_1(t), g_2(t), \dots, g_n(t)], \tag{2.2}$$

where m is the number of segments in σ.

To keep track of which is the first and second argument in $a_s(\cdot, \cdot)$ we use a directed graph c with the convention $s = \overline{(i_1, i_2)}$.

To begin with we shall take a *nonparametric approach*, letting all the mr^2 entries in the a_s matrices be free. Of course we can add a constant k to all the entries if we at the same time multiply Z by e^k. We do not have full identifiability. So we ask that $Z = 1$, introducing of course relations between the a's.

The lack of identifiability is, however, more serious than so. We shall devote the next section to clarifying this issue before we begin the construction of estimates. Identifiability as such could be inferred from results in Besag (1975), but we need the following formal developments for Theorems 2–6 and for studying another estimation method.

2.1. Let us consider an example, see Figure 10, in which we show a directed graph σ with $n = 6$ sites and $m = 7$ segments and G, r arbitrary but finite. The arities ω_i are also shown. We assume the graph to be directed in order to distinguish between the endpoints of any edge. Compare with examples in Sections 1.2.1 and 1.2.3 in which it is also seen that we must be prepared to let

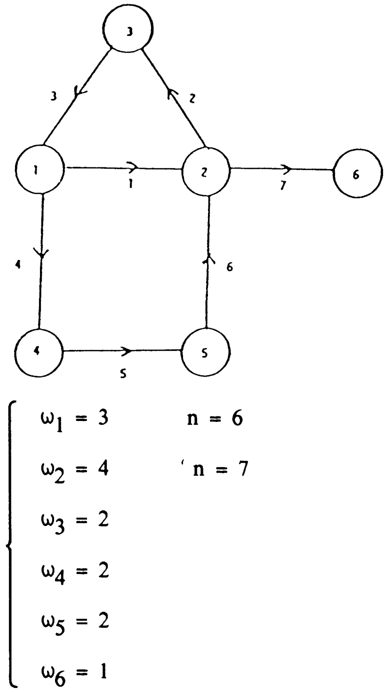

FIG. 10. *Configuration diagram.*

in-arities differ from out-arities. It is easy to see that the a_s entries can be modified in many ways without changing the measure P. We are actually dealing with exponential families and the trouble is caused by singularity of a matrix; see Barndorff-Nielsen (1978).

We must therefore restrict our parameter space by replacing it by a subspace. When we do this we must ask that (a) the new parameters are uniquely determined by a P of the form (2.1) and (b) that the family $\{P\}$ is not reduced.

This can be done in many ways; the following one is both theoretically attractive and computationally convenient. We do it by the following *cutting system*: *Cut the segments* (at one side, at both sides or not at all) so that

1. *exactly one segment remains uncut*;
2. *for each site there is exactly one joining segment that is not cut close to the site*.

In our example we show one such cut system in Figure 11. Segment 5 is the uncut one. One can prove that this can be done for any finite connected graph, usually in many ways.

We can use the following:

CUTTING ALGORITHM.

0. Sets S and A are empty.
1. Choose one segment in the graph and leave it uncut. Add the segment to set S and one of the sites to set A.
2. Choose a segment in the graph that connects to *at least* one site in A.
   (a) If only one of the two sites that are connected by the segment is in A, then cut the segment close to that site. Add the segment to set S and add the other site to set A.
   (b) Otherwise cut the segment twice. Add the segment to the set S.
3. Repeat Step 2 until all segments have been added to S.

With such a cutting system we can establish identifiability. Select one generator (arbitrarily) to be denoted by 0.

THEOREM 2. *Standardize the acceptors by*

$$\begin{aligned} a_s(g,0) &= 0, \\ &= g, \quad \textit{if } s = \left(\overrightarrow{i, j}\right) \textit{ is cut close to } i, \\ a_s(0,g) &= 0, \\ &= g, \quad \textit{if } s \textit{ is cut close to } j. \end{aligned} \tag{2.3}$$

Then the remaining a_s entries are uniquely determined by P.

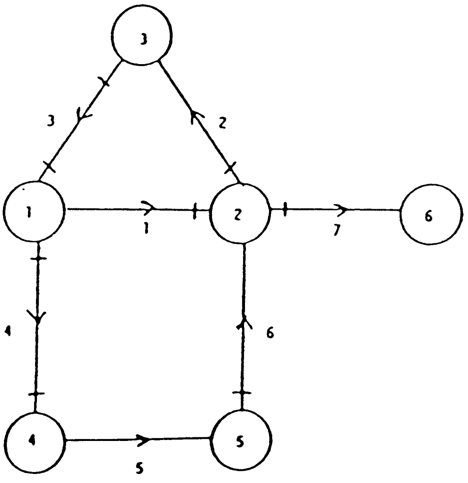

FIG. 11. *Cut configuration.*

We can be more specific. Introduce the difference operators for some function $f = f(g_1, g_2, \dots, g_i, \dots, g_n)$,

$$\begin{aligned} \Delta_i f &= f(g_1, g_2, \dots, g_{i-1}, g_i, g_{i+1}, \dots, g_n) \\ &\quad -f(g_1, g_2, \dots, g_{i-1}, 0, g_{i+1}, \dots, g_n), \\ (f)_i &= f(0, 0, \dots, 0, g_i, 0, \dots, 0). \end{aligned} \tag{2.4}$$

We can then derive the following explicit representations for the acceptor entries in terms of the log probabilities $\bar{p}(c) = \log P(c)$.

THEOREM 3. *With a cutting system as defined we have*

$$a_s(g_i, g_j) = \begin{cases} \Delta_i \Delta_j \bar{p} + (\Delta_j \bar{p})_j, \\ \qquad \textit{if } s = \left(\overrightarrow{i, j}\right) \textit{ is cut close to } i, \\ \Delta_i \Delta_j \bar{p} + (\Delta_i \bar{p})_i, \\ \qquad \textit{if } s \textit{ is cut close to } j, \\ \Delta_i \Delta_j \bar{p}, \\ \qquad \textit{if } s \textit{ is cut twice}, \\ \Delta_i \Delta_j \bar{p} + (\Delta_i \bar{p})_i + (\Delta_j \bar{p})_j + \bar{p}(0, 0, \dots, 0), \\ \qquad \textit{if } s \textit{ is uncut}. \end{cases} \tag{2.5}$$

The important *homogeneous acceptor case*, when all A_s are the same, has also been solved. We can state

THEOREM 4. *In the homogeneous acceptor case identifiability is achieved*

(a) *if for each site i the in-arity equals the out-arity,* $\omega^i_{\text{in}} = \omega^i_{\text{out}}$, *use the restriction* $a(0, g) = 0$, $g \in G$;
(b) *otherwise no restriction is needed.*

2.2. Now the estimation problem is correctly posed and we can ask for an estimate a^* of the whole set of free acceptor entries organized as a vector a. The likelihood function for our sample is then

$$L_N(a) = Z^{-N}(a) \prod_{t=1}^{N} \exp\left\{ \sum_{s=1}^{m} a_s\left[g_{i_1}(t), g_{i_2}(t)\right] \right\}. \tag{2.6}$$

To use *maximum likelihood* estimation, which is well advised in principle for an exponential family situation, is *not computationally feasible* because of the difficulty of evaluating $Z(a)$.

Instead we shall use a modification of the ingenious *pseudolikelihood* method introduced by Besag (1974); see also Possolo (1985), which contains many references to this subject. Introduce the pseudolikelihood

$$\mathrm{PL}_N(a) = \prod_{t=1}^{N} \prod_{i=1}^{n} P^a\left[g_i(t) | \mathrm{env}_i(t)\right], \tag{2.7}$$

where

$$P^a\left[g_i(t) | \mathrm{env}_i(t)\right] = P^a\left[g_i(t) \mid \text{all neighbors of } g_i \text{ in } c(t)\right]. \tag{2.8}$$

This function does not involve the troublesome partition function $Z(a)$ since we can write

$$P^a[g | \mathrm{env}_i] = \frac{1}{Z^i_{\mathrm{env}}(a)} \exp \sum_{j=1}^{\omega_i} a_{s_j}\left(g, \mathrm{env}_i^j\right) \tag{2.9}$$

with

$$\mathrm{env}_i^j = j\text{th neighbor of site } i \tag{2.10}$$

and

$$Z^i_{\mathrm{env}}(a) = \sum_{g \in G} \exp \sum_{j=1}^{\omega_i} a_{s_j}\left(g, \mathrm{env}_i^j\right). \tag{2.11}$$

We now define the *pseudolikehood* estimate a_N^* by solving

$$\mathrm{PL}_N(a^*) = \max \mathrm{PL}_N(a), \tag{2.12}$$

where, as before, the parameter space has been restricted according to a cutting system for the connector σ. It can be shown that as $N \to \infty$ the probability that (2.12) has a unique solution tends to 1 and, furthermore,

THEOREM 5. *The pseudolikelihood estimate is consistent.*

2.3. We can now go ahead and study the asymptotics of the pseudolikelihood estimate. We derive the covariance matrix for the asymptotically normal distribution of a^* and get

THEOREM 6. *The pseudolikelihood function provides an estimate of a that is asymptotically normal with mean* a_0 (*the true value of the acceptor functions*) *and with covariance matrix*

$$\frac{1}{N} K^{-1} S K^{-1},$$

where

$$K = -E\left[\sum_{i=1}^{n} \nabla_a^2 \log P_i[g|\mathrm{env}]\right]_0,$$

$$S = E\left[\left(\sum_{i=1}^{n} \nabla_a \log P_i[g|\mathrm{env}]\right)\left(\sum_{i=1}^{n} \nabla_a \log P_i[g|\mathrm{env}]\right)^T\right]_0$$

and $[\cdot]_0$ *indicates that the expression is to be evaluated at the point* a_0.

2.4. Software has been developed for cutting the connector σ and computing the pseudolikelihood estimate. The latter is done by exploiting the fact that $\mathrm{PL}_N(a)$ is a logarithmically concave function of a. This facilitates the search for the maximum in parameter space and guarantees convergence of the search algorithm.

A systematic series of computer experiments has been designed and carried out. In order to be able to compute efficiencies exactly we have chosen small graphs but the applicability and feasibility of the software is not restricted to small n. For larger graphs we recommend partitioning σ,

$$\sigma = \sigma_1, \sigma_2, \ldots,$$

and applying the program to each σ_k separately, iterating this a number of times. Convergence follows again using the concavity of $\log \mathrm{PL}_N$.

In the cases studied the obtained values for the asymptotic efficiencies (relative to maximum likelihood) fell in the range 90–100% in all cases.

One of the purposes of this computer experiment was to investigate a "local difference" estimator, using the equations in (2.5) to estimate the a_s-values. To do this one has to estimate the $\bar{p}$ values on the right-hand side of this equation, which requires some care since many of the corresponding p values can be zero or close to zero. This is very fast computationally and gave surprisingly good results [Grenander and Osborn (1985)] if the graph was of small or moderate size. The first versions of the software required much CPU time but improved versions achieved higher speed. Computing speed is not critical here, however, if estimation is going to be done once and for all off line. This is in contrast to stochastic relaxation which is computed for each observed image, when we are sampling from the same image prior.

The asymptotic sampling errors have also been used for the following purpose. When a *complex system is going to be predicted by our parallel logic*, say by predicting the g_i value, given values for $g_{i_1}, g_{i_2}, \dots, g_{i_l}$, we can first use Theorem 1 to find what sites, if any, can be neglected among the observed ones. Then we simulate

$$P^{a^*} = P^{a^*}\left(g_i = g | g_{i_1}, g_{i_2}, \dots, g_{i_l}\right),$$

which will normally differ from

$$P^a = P^a\left(g_i = g | g_{i_1}, g_{i_2}, \dots, g_{i_l}\right).$$

Our results allow us to make approximate statements about the error

$$E_a\left(P^a - P^{a^*}\right)^2.$$

The results in Section 2 are given with proofs and numerical results in Grenander and Osborn (1985).

3. Limit theorems in pattern theory. In order to derive approximations to the Markov process measures induced by the structure formulas we shall consider the following types of *limit problems in metric pattern theory*:

1. Asymptotics of P as ε or $T \downarrow 0$, fixed σ.
2. Asymptotics of P as $n = |\sigma| \to \infty$, fixed ε, T.

The case of greatest interest is, however, the *mixed limit problem*:

3. Asymptotics of P as $\varepsilon, T \downarrow 0$ and $n = |\sigma| \to \infty$.

It is considerably more difficult than the two first limit problems and we shall postpone discussing it until Section 3.3.

3.1. In *the first limit problem* we keep everything fixed, including the connector σ, except temperature T or the scale factor ε. When they are made to tend to zero the regularity $\langle G, \sigma, A(\varepsilon)\rangle$ or $\langle G, \sigma, A(T)\rangle$ will become more rigid and approach the *frozen patterns*.

Let us write the exponent in the structure formula (1.2), absorbing the Q factors into the first product, as

$$-\frac{1}{T}H(c), \tag{3.1}$$

where H corresponds to the Hamiltonian energy term in statistical mechanics. Then large values of the probabilities $P(c)$ correspond to low values of the energy $H(c)$. Therefore the set

$$M = \left\{c|H(c) = \min_{c} H(c)\right\} \subset C \tag{3.2}$$

of *minimum energy configurations* can be expected to appear in the solutions of the first limit problem. Without loss of generality we assume that the minimum is zero.

For simplicity we shall assume that H is continuous and the minimum in (3.2) is attained.

The results in Section 3.1 are due to Hwang (1980, 1981).

3.1.1. The simplest case is of course when the generator space is finite when we can state

THEOREM 7. *If* $|G| < \infty$ *we have*

$$\lim_{T \downarrow 0} P_T(c) = \begin{cases} \dfrac{1}{|M|}, & \textit{if } c \in M, \\ 0, & \textit{else}. \end{cases} \tag{3.3}$$

In other words the limit measure is uniform on the set of minimal energy configurations which is intuitive.

3.1.2. We turn now to the case $|G| = \infty$ and shall assume that $G = R$ so that c can be embedded in a vector in R^n. If G is some other finite-dimensional Euclidean space, the following holds with obvious modifications.

The densities $p(c) = f_T(c)$ in the structure formula are viewed as Radon–Nikodym derivatives with respect to some fixed probability measure m:

$$\frac{P_T(dc)}{m(dc)} = f_T(c).$$

Assume

$$m\{c|H(c) < a\} > 0 \quad \text{for } a > 0. \tag{3.4}$$

THEOREM 8. *If* (3.4) *holds and* $m = m(M) > 0$ (*where m is not the same as the earlier m*) *we have the limiting probability measure given by*

$$\lim_{T \downarrow 0} P_T(A) = \frac{1}{m} m(A \cap M). \tag{3.5}$$

Again the limit is uniform on the set of minimum energy configurations.

The case $m = 0$ is a little harder. Assume first that M is finite $= \{x_1, x_2, \dots, x_s\}$, that there exists an $\varepsilon > 0$ such that $\{x | H(x) \leq \varepsilon\}$ is compact and that

$$H \in C^3(\mathscr{R}^n), \quad \frac{m(dx)}{\mu(dx)} = f(x) \text{ is continuous}, \quad \mu = \text{Lebesgue measure}. \tag{3.6}$$

Then the following describes the limiting measure.

THEOREM 9. *Under the given conditions and assuming that the Hessian* $H''(x_i)$ *of* $H(x)$ *is nonsingular for all i and that for some k* $f(x_k) > 0$, *then the limiting measure is concentrated on M with the mass at* x_i:

$$\frac{f(x_i)\left[\det H''(x_i)\right]^{-1/2}}{\Sigma_{j=1}^{s} f(x_j)\left[\det H''(x_j)\right]^{-1/2}}. \tag{3.7}$$

If M is not a finite set things become more complicated and the conditions needed will become somewhat elaborate. Without stating the conditions in detail [they can be found in Grenander (1981), pages 220–230], let us assume that M consists of a finite number of compact smooth manifolds and let N consist of the highest-dimensional manifolds. Then one can prove

THEOREM 10. *The limiting probability measure is concentrated on N with a density proportional to*

$$f(u)\left[\det\left(\frac{\partial^2 H(u)}{\partial t^2}\right)\right]^{-1/2}, \tag{3.8}$$

where u is a coordinate vector on N and the derivatives with respect to t mean differentiation in normal directions at $u \in N$.

The density in (3.8) is taken with respect to a certain intrinsic measure on N.

In the special case when H is a linear-quadratic function so that the P_T measures are Gaussian, we can use special methods. Indeed, let us consider

$$H(x) = x^T F x - k^T x, \qquad k \in \mathscr{R}^n. \tag{3.9}$$

THEOREM 11. *In the Gaussian case* (3.9) *a limiting measure exists for* $T \downarrow 0$ *if and only if F is nonnegative definite and k is in the range of F. Then the limiting Gaussian measure has as covariance operator the projection down to the null space of F and the mean m is given by* $k = Fm$.

3.2. *In the second limit problem* we keep T or ε fixed so that we need not indicate them in our formulas. It is the connector σ that is made large in some sense, so that $n \to \infty$.

3.2.1. Let σ be a linear graph and consider the frequency function with respect to Lebesgue measure in $\mathscr{R}^n$,

$$p_n(c) = \frac{1}{Z_n} \prod_{i=1}^{n-1} A(g_{i+1}, g_i) Q(g_i), \tag{3.10}$$

where G is a compact interval on R, e.g., $[-1,1]$. Consider the marginal distribution of g_i, $i = [\alpha n]$, $0 < \alpha < 1$, as $n \to \infty$. The reason we ask i to behave like this is that we want to avoid boundary effects for i close to 1 or n. One can prove convergence for such marginal distributions.

The density (3.10) defines a Markov chain with $(-1,1)$ as a state space. Note, however, that:

1. The transition probabilities are not time homogeneous; they depend upon i.
2. The transition probabilities will usually depend on n.
3. When n increases, P_{n_1} will not necessarily be a marginal measure of P_{n_2}, $n_1 < n_2$.

These circumstances make for some technical difficulties in the proof of the following result. Assume that A is symmetric, continuous and strictly positive and introduce the integral operator

$$(Tf)(x) = \int A(x, y) Q(y) f(y)\, dy.$$

The kernel is not symmetric but can be symmetrized by the transformation $f \to f\sqrt{Q}$. One can then derive the following answer to the limit problem of the second type.

THEOREM 12. *The marginal distribution defined above converges weakly as $n \to \infty$ to a probability measure with a density*

$$\frac{\phi_1^2(x)}{\int \phi_1^2(x)\, dx}, \tag{3.11}$$

where ϕ_1 is an eigenfunction associated with the smallest eigenvalue of T.

This is due to Plumeri (1981).

3.2.2. To extend this to other connectors besides the linear ones has not yet succeeded for conditions as general as above. In the Gaussian case, however, it has been done and we first illustrate the sort of results that are possible by again looking at a simple graph.

Let σ be a generalized cyclic graph with n sites (for the linear graph the same results can be obtained) and let the bond structure be as exemplified in Figure

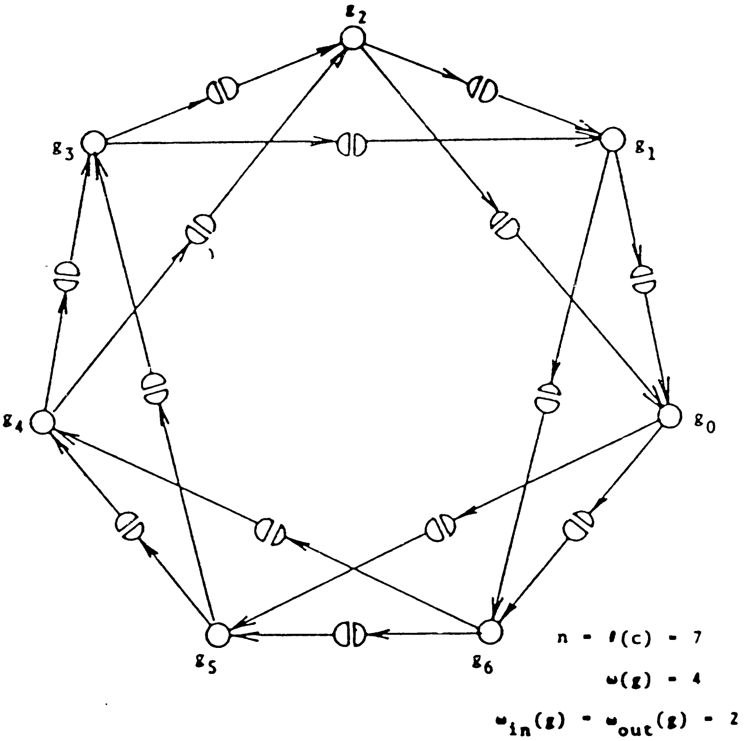

FIG. 12. *Connector graph.*

12, each site having $2p$ neighbors, $p = 2$ in the figure.

Let the Q factors be Gaussian, mean value vector zero and

$$g_i = \begin{pmatrix} X_i \\ Y_i \end{pmatrix}, \qquad X_i \in \mathscr{R}^p, \qquad Y_i \in \mathscr{R}^p, \tag{3.12}$$

where X_i represents the bonds sent in one direction from g_ι and Y_i the bonds in the other direction. Let us use rigid regularity with the bond relation ρ meaning equality and the kth bond of X_i connected to, say, the kth bond of Y_{i+a_k}.

Let the covariance matrix of Q be given by H^{-1} where H is the partitioned matrix in the exponent

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{12}^T & H_{22} \end{pmatrix}, \quad \text{all } H_{kl} \text{ are } p \times p. \tag{3.13}$$

Introduce the fundamental circulant $n \times n$ matrix $\pi_n = (\delta_{ii+1})$ with addition modulo n and the matrices

$$E_{k,p} = \operatorname{diag}(0, 0, \ldots 1, \ldots, 0) \quad \text{with the 1 in the } k\text{th place.} \tag{3.14}$$

To find the limiting measure as $n \to \infty$ we use the following construction. Consider the sum of $2p \times 2p$ matrices

$$\begin{aligned} &\Pi_n^0 \otimes H_{11} + \sum_{k=1}^{p} \Pi_n^{-a_k} \otimes H_{12}E_{k,p} + \sum_{k=1}^{p} \Pi_n^{a_k} \otimes E_{k,p}H_{12}^T \\ &\quad + \sum_{k=1}^{p} \Pi_n^{a_k} \otimes E_{k,p}H_{22}) \left(\sum_{l=1}^{p} \Pi_n^{a_k} \otimes E_{k,p} \right), \end{aligned} \tag{3.15}$$

where $A \otimes B$ means the matrix $\left(\begin{smallmatrix} A & 0 \\ 0 & B \end{smallmatrix}\right)$.

Collecting terms in (3.15) the sum can be expressed as a sum of the form

$$\sum_{k=-q}^{q} \Pi_n^k \otimes C_k \tag{3.16}$$

with new matrices C_k of size $p \times p$ and where q is some natural number..

As $n \to \infty$ what limiting measure do we get for the vector valued stochastic process $(X_i;\ i = 0, 1, 2, \dots, n-1)$? The answer is given by

THEOREM 13. *The limiting Gaussian process is stationary with mean zero and spectral density matrix*

$$f(\lambda) = \left(\sum_{k=-q}^{q} C_k e^{-ik\lambda} \right)^{-1}. \tag{3.17}$$

Many other graphs of lattice type with arities greater than 1 can be solved in a similar way, so that the *spectral density matrix* can be obtained by purely algebraic manipulations like the ones in (3.15)–(3.17). These results are due to P. Thrift; more theorems and proofs can be found in Grenander [(1981), pages 252–288].

3.3. The results in Theorem 7.13 illustrate what sort of limit theorems can be obtained when either σ or T and ε is fixed. The practically most important case is when both tend to their respective limit since this is when stochastic relaxation is most time consuming and we need approximative procedures based on analytical results.

Progress in that case has been slow and the only result obtained until recently was the one to be presented in Section 3.3.1. It was clear that a different technique was needed that was less dependent upon the particular form of the connector σ. Such a method has been suggested and used recently and will be discussed in Section 3.3.2.

3.3.1. To study *the mixed limit problem*, say that σ is a linear graph so that our probability measure is given by the density in (3.10). Since we are now going to vary ε it should be included in the formula and we write

$$(3.18)\qquad p(c) = p_{n,\varepsilon}(c) = \frac{1}{Z_{n,\varepsilon}} \prod_{i=0}^{n-2} A\left(\frac{g_{i+1} - g_i}{\varepsilon}\right) Q(g_i)', \qquad g_i \in \mathscr{R}.$$

We shall assume that $Q \in C_2$ with a unique global maximum, say at $g = 0$, and with a second derivative $q = -Q''(0) > 0$. The choice of the acceptor function $A(x)$ is believed not to be crucial as long as it decreases fast enough as $|x| \to \infty$ to avoid long range dependence. In this section we shall use a rectangular window $A(x) = 1$ if $|x| \leq 1$, 0 else, but the results are of wider validity.

If we let $n \to \infty$ first, with ε constant for the moment, we know that the limit can be expressed via the first eigenfunction ϕ_1 of the integral equation

$$\lambda\phi(x) = \int A\left(\frac{x - y}{\varepsilon}\right) Q(y)\phi(y)\, dy$$

because of Theorem 12. Now let $\varepsilon \downarrow 0$. How does ϕ_ε behave? Using reasoning similar to that underlying Theorem 8 it is easy to show that ϕ_1 will contract to a one-point distribution with the mass at the maximum of Q. We need more detailed knowledge, however, and a computer experiment was carried out in order to guide us.

For a variety of Q functions the eigenfunction ϕ_1 was computed and plotted for a sequence of small ε values. The graphs were striking—they look very much like Gaussian densities!

But why? Standard perturbation calculus for the operators T_ε will not be enough since the limiting operator is not compact: We are dealing with a *singular perturbation problem*. The solution is given by

THEOREM 14. *The normalized density $\sqrt{\varepsilon}\,\phi_\varepsilon(\sqrt{\varepsilon}\, x)$ tends weakly, as $\varepsilon \downarrow 0$, to a Gaussian limit with mean zero and variance $1/\sqrt{3q}$.*

The proof, which is long and technical, is given in Chow and Grenander (1981). One can also show that the limit remains the same when ε and n tend to their limits under a condition that says that n increases slowly enough compared to the rate by which ε tends to zero.

3.3.2. The proof of Theorem 14 does not shed much light on why the limit is Gaussian, it is too calculating and not intuitive enough. Also it does not seem to be extendable to other connectors, for example, the cyclic one that appears in some useful random geometries describing the shape of random objects; see Section 1.8.3. One reason for this is that one has not succeeded in extending Theorem 12 [some results for σ = tree can be found in Plumeri (1981)], which underlies Theorem 14, to general σ.

Instead a completely different analytical technique has been developed in Grenander and Sethuraman (1985). Here we try to make statements not just about the asymptotics of marginal distributions of some g_i but about the whole measure.

Say that σ is a cyclic chain so that (3.18) holds if modified by including a factor $A((g_{n-1} - g_0)/\varepsilon)$ on the right-hand side. For analytical convenience we choose $A(x)$ to be of the form $\exp(-cx^2)$ and introduce a stochastic process $C_n(t)$, $0 \le t \le 1$, by defining

$$C_n\left(\frac{i}{n}\right) = g_i, \quad C \text{ linear between } \frac{i}{n} \text{ and } \frac{i+1}{n}. \tag{3.19}$$

We think of $[0,1)$ as a parametrization of a circle M of radius 1. Then under some assumptions on Q we have

THEOREM 15. *The process $c_n(t)$ converges weakly, for $\varepsilon = 1/n$ and $n \to \infty$, to a process $c(t)$ which is stationary (on the circle M), Gaussian, and Markovian on M with the covariance function*

$$R(s,t) = \frac{\cosh\left[(s-t-1/2)/\sqrt{q}\right.}{2\sqrt{q}\,\sinh(\sqrt{q}/2)}. \tag{3.20}$$

A *computational consequence* of Theorem 15 is that since $c(t)$ can be simulated directly by the representation

$$c(t) = \frac{e^{\sqrt{q}/2}}{2\sinh(\sqrt{q}/2)} \int_0^1 e^{-\sqrt{q}[(t-s),\,\mathrm{mod}\,1]} w(ds), \tag{3.21}$$

where w is the standard Wiener process, we can avoid stochastic relaxation and obtain an extremely fast simulation procedure for the approximating process.

But, more importantly, the proof technique used for Theorem 15 *does not rely on the special geometry of the connector*. Therefore it seems possible to extend the result to many types of lattice graphs and this is being done at present. Without going into this, it should be mentioned that for some such connectors the limiting measure has to be associated with stochastic processes whose "sample functions" are not ordinary functions, but Schwartz distributions.

4. Open problems. The theoretical developments described above leave a number of questions unsettled. Let us briefly mention a few of the major ones.

4.1. In a typical pattern theoretic setup some of the generators are *unobservable in principle*: We speak of invisible sites. For example, in Figure 3 the sites in the upper level cannot be seen by the observer.

To determine the acceptor values empirically using a set of pure images, we cannot use the procedure of Section 2.2 without modification. Attempts have been made to combine the pseudolikelihood approach with the EM method, but so far this has been done only numerically and we have no theoretical support for it.

For rigid regularity one can often determine the configuration from the image in a unique manner and the above problem does not arise. Although this will take care of many situations it is clearly desirable to look more carefully into the theoretical aspect. To do this, one should first of all seek a standardization of the acceptors so that they become identifiable. This question is similar but distinct from the one answered in Section 2.1 and we have no answer to it at present.

4.2. If only the deformed, but not the pure, images can be observed we get an additional difficulty—*noisy data*. This seems less formidable than the one associated with hidden sites. Indeed, the probability measure for the I's is, for most deformation mechanisms, identifiable from the knowledge of the probability measure of $I^{\mathscr{D}}$. Therefore there is good hope to derive estimators of I given $I^{\mathscr{D}}$ and thus eliminate the difficulty, but this has not yet been done.

4.3. If the generator space G has large cardinality or is infinite, for example $G = \mathbb{R}$, it must be discretized before our estimation results can be applied. The *level of discretization* must of course be related to the sample size N: Large N will allow finer discretization.

It is natural to do this by a sieve when choosing the bond value space B and let it have a small number m of elements. As N increases we can let m increase (slowly!), or, equivalently, make the mesh size $\mu = 1/m$ of the sieve tend to zero.

4.4. In Theorem 4 we showed how to standardize acceptors in the homogeneous case. This should be generalized to connector graphs $\sigma = \bigcup_k \sigma_k$, where the acceptors are the same, A_k, over each subgraph σ_k. It is not known at present how to construct an appropriate standardization.

4.5. We have assumed throughout that the connector is known. Often this is the case and σ may be the structure of a scientific doctrine, for example. *If the connector σ is unknown*, however, we need estimation methods to determine σ. The only such procedure available for $G = \mathbb{R}$, seems to be to compute $(R^*)^{-1}$, where R^* is the empirical correlation matrix for the g_i's, and look for entries with large absolute values. This seems to work to a limited extent in cases when all regressions are approximately linear, but we have no computationally feasible method for general applicability.

4.6. Consider stochastic relaxation with a deterministic, periodic sweep strategy; see Section 1.6. Let one full sweep constitute one unit of time.

How fast do the measures over C converge to the equilibrium measure P? The *Markov chain in time*, with C as the state space, has some transition probability matrix

$$M = \big(P(c \to c' \text{ after one sweep});\ c, c' \in C\big). \tag{4.1}$$

The next largest (in absolute value) eigenvalue λ_2 of M determines the speed of convergence. The smaller λ_2 is the faster is the convergence.

In particular, does $\lambda_2 \to 1$ as $n \to \infty$ and if so, how fast? This is the so-called *λ_2 problem*. It could be that more and more full sweeps are required as the connector graphs are made large. Numerical experiments indicate that, for the regularities studied, this does not happen, but there is little theoretical support for this conjecture at present.

4.7. If one could get a bound or an approximation for λ_2 this would give us a tool for choosing a *good sweep strategy*. We would select one for which λ_2 was small.

A remarkable numerical result was obtained by D. E. McClure, who computed λ_2 for connectors small enough for direct computation to be possible. For each of those connectors, for $|G| = 2$, and for a number of deterministic, periodic sweep strategies he got the same λ_2. This was done in double precision, about 16 decimal digits. This surprising observation has not been explained theoretically.

REFERENCES

BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families in Statistical Theory.* Wiley, New York.

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **36** 192–236.

BESAG, J. (1975). Statistical analysis of non-lattice data. *The Statistician* **24** 179–195.

CHOW, Y.-S. and GRENANDER, U. (1985). A singular perturbation problem. *J. Integral Equations* **9** 63–73.

GEMAN, D. and GEMAN, S. (1983). Parameter estimation for some Markov random fields. Brown Univ. Complex Systems No. 11.

GEMAN, D. and GEMAN, S. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intelligence* **6** 721–741.

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

GRENANDER, U. (1976, 1978, 1981a). *Lectures in Pattern Theory* **1–3**. Springer, Berlin.

GRENANDER, U. (1981b). *Abstract Inference*. Academic, New York.

GRENANDER, U. (1983). Tutorial in pattern theory. Unpublished.

GRENANDER, U. and KEENAN, D. M. (1988). HANDS. A pattern theoretic approach to processing biological images. Unpublished.

GRENANDER, U. and OSBORN, B. (1985). Estimation problems in pattern theory. Unpublished.

GRENANDER, U. and SETHURAMAN, J. (1985). Limit theorems in metric pattern theory. Unpublished.

HARDING, E. F. and KENDALL, D. K., eds. (1974). *Stochastic Geometry.* Wiley, New York.

HWANG, C.-R. (1980). Laplace's method revisited: Weak convergence of probability measures. *Ann. Probab.* **8** 1177–1182.

HWANG, C.-R. (1981). A generalization of Laplace's method. *Proc. Amer. Math. Soc.* **82** 446–450.

KINDERMAN, R. and SNELL, J. L. (1980). *Markov Random Fields and Their Applications*. Amer. Math. Soc., Providence, R.I.

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

PLUMERI, C. (1981). Probability measure on regular structures. Ph.D. dissertation, Div. Applied Mathematics, Brown Univ.

POSSOLO, A. (1985). Estimation of binary Markov random fields. Unpublished.

THRIFT, P. (1979). Autoregression in homogeneous Gaussian configurations. Ph.D. dissertation, Div. Applied Mathematics, Brown Univ.

DIVISION OF APPLIED MATHEMATICS
BROWN UNIVERSITY
PROVIDENCE, RHODE ISLAND 02912