RANDOM FIELDS AND INVERSE PROBLEMS IN IMAGING*

D. GEMAN

* This work was partially supported by the National Science Foundation under grant DMS-8813699 and by the Office of Naval Research under contract N00014-88-K-0289

TABLE DES MATIERES

D. GEMAN : "RANDOM FIELDS AND INVERSE PROBLEMS IN IMAGING"

1 Introduction

Image Analysis. A digital image is a matrix of positive integers which represents a pattern of radiant energy emitted by objects in space. The image *synthesis* (or *direct*) problem is to determine the digital image given the scene geometry, sources of illumination, and so forth, and is in the domain of optics, sensor modeling, and computer graphics. These lectures concern image *analysis*, a general term encompassing problems in image processing, such as removing the effects of blur and noise, and those in computer vision, which involve geometric and semantical scene descriptions. Whereas specific goals are application-dependent, one can still isolate many generic issues and sub-problems, all involved with converting information which is *implicit* in the recorded digital image to *explicit* descriptions of the physical world, a process sometimes characterized as "inverse optics"[111].

Mathematically, problems in image analysis range from ill-conditioned to ill-posed. This is due to the information loss in passing from the continuous dimensions of the physical world to the sampled and quantized image values. Take, for example, the simplest case of optical blurring $y = \mathcal{K}x$, in which the original and blurred images are represented as vectors $x \in \mathbf{R}^N$ and $y \in \mathbf{R}^M$ and $\mathcal{K}$ is an $M \times N$ matrix representing the point spread function. Due to the nature of image blurring and data acquisition, this linear system is typically underdetermined ($M < N$) and two estimates of x, both faithful to the data y, may be very far apart, both visually and as vectors. Moreover, this "exact data" case is an idealization; in reality there is noise, at least measurement error due to the quantization of the data to discrete values. Consequently, even if $\mathcal{K}$ were invertible, it is usually nearly singular, and the inverse problem is then ill-conditioned (unstable) due to existence of noise. Similar situations are encountered in particle scattering, spectroscopy, and radio astronomy.

This course will focus on inverse problems in "low-level" image analysis, for example, restoring images degraded by blur and noise, and tomographic reconstruction from photon counts, and in "middle level" image analysis, such as texture segmentation and boundary detection. Other problems in this class include the synthesis of the reflectance map of natural objects (such as textures or landscapes), the computation of optical flow from motion sequences, extracting range measurements from multiple-angle views (stereo-matching), and estimating shape (e.g., surface orientation) from texture, motion, or shading.

Biological Solution. Animals solve these problems in an apparently effortless way. This seemingly miraculous ability is attributed to several factors, not the least of which is the large proportion of the brain devoted to vision. We are able to integrate cues from many sources, including binocular stereo, motion, and color. In addition, we exploit a priori expectations, specific scene knowledge, and contextual clues to reduce, or even remove, ambiguity and perceive the world "correctly". Except for rare cases, automated visual systems are vastly inferior to natural ones, due somewhat to a lack of raw processing power (or suitably parallel computation) but much more importantly to the inability to integrate information and place appropriate constraints. Put differently, the weak link in computer vision is the software, not the hardware.

Stochastic Constraints. A current trend is "stochastic regularization", in which random field models are employed to provide constraints and fuse information. In real scenes, nearby locations typically have similar intensities; boundaries are usually smooth

and persistent; textures, although sometimes locally random, define spatially homogeneous regions; and objects, such as roads and leaves, have preferred relations and orientations. These "regularities" are rarely deterministic; rather, they describe correlations and likelihoods, and this kind of generic or a priori knowledge can be captured mathematically and exploited in a stochastic framework to make inferences. Properly conceived, such models impose severe, but appropriate, restrictions on the set of plausible interpretations. What is needed then, as Grenander [68] emphasizes, is a "general pattern theory" in which these "prior" constraints are precisely formulated and systematically combined with empirical information in order to make inferences about spatial patterns and image attributes. Ideally, as noted in [68], such a theory would also provide feasible algorithms and measures of optimality and performance much in the same way that the theory of stationary stochastic processes provides a natural mathematical framework for *linear* estimation. This at least is the program; what already exists, and is the purpose of these lectures to illustrate, is a serious beginning, involving a rich variety of models and image problems. It remains to be seen whether this methodology can be simplified to the level of everyday use (say in processing medical or satellite data) or for that matter advanced to the level of more complex problems, such as those involving recognition and interpretation.

Bayesian Inference. There are usually several image attributes of interest, which, upon digitization, may be regarded as (finite) two-dimensional arrays. One of these is a regular lattice of quantized brightness values corresponding to the "true" distribution of radiant energy; the lattice elements or "pixels" correspond to small patches in the (idealized) scene. These values may or may not be part of the data, i.e., directly observed. The other arrays represent problem-specific image attributes, for instance classification or boundary labels. This collection of arrays, denoted, say, by X, is modeled as a discrete-parameter stochastic process indexed by the vertices of a graph, $\mathcal{G}$. The vertices or "sites" of $\mathcal{G}$ serve simply to index the process whereas the edges or "bonds" of $\mathcal{G}$ capture the *interactions* among the individual random variables. For computational reasons (see below) these graphs are usually *sparse* in the sense that the neighborhoods are small compared to the graph size; this condition is also rather natural for modeling spatial phenomena. Specifically, the process X is a Markov random field with respect to $\mathcal{G}$, and hence its probability law $\Pi(x)$ is a Gibbs distribution over $\mathcal{G}$, sometimes referred to as the "prior distribution" for X. Section 3 is devoted to the relevant theory and a number of examples extracted from the literature on imaging.

Now let Y denote the observation process, so that, by definition, the *data* is $Y = y$, and let $\Pi(y|x)$ be the conditional distribution of Y given X. (Actually, information may be available from multiple views, sensors, or wavelengths, but we shall imagine a single observation process.) The transformation from X to Y is often nonlinear and at least partially random, and may involve, for example, optical blurring, quantum noise, obscurations (i.e. missing observations), or a Radon transform. In the simplest case of fully observed and uncorrupted intensities, the distribution $\Pi(y|x)$ is degenerate since one component of X is then Y itself.

These two distributions, $\Pi(x)$ and $\Pi(y|x)$, obviously determine the *joint distribution* of X and Y, and in particular the *posterior distribution* of X given Y, denoted $\Pi(x|y)$. Under reasonable conditions on the degradation mechanism (the blur extent, noise correlations, etc.), this is another Gibbs distribution over a graph, say $\mathcal{G}^P$, which is in general

different from $\mathcal{G}$, but still sparse. Equivalently, conditional on the data, the process X may be regarded as another Markov random field with limited direct interactions (although possibly long-range correlations). Finally, the actual estimate of X , say $\hat{x} = \hat{x}(y)$, is then defined in terms of $\Pi(x|y)$. For example, $\hat{x}$ might be the mean or mode of $\Pi(x|y)$, the Bayes estimates corresponding to squared-error and zero-one loss respectively. Or the object of interest may simply be a sample from $\Pi(x|y)$, as in texture synthesis.

Markov Random Fields. Bayesian inference is not new in image analysis (see e.g. [71],[83],[107],[117]). However, there has been considerable recent interest, particularly in prior models based on discrete Markov random fields over graphs. For instance, natural textures such as grass or wood are synthesized by sample realizations by adjusting parameters corresponding to interactions at different scales and orientations. In image restoration and reconstruction the "true" array of brightness values is again conceived as the sample realization of a Markov random field (over a lattice-based graph), but the object here is to recover this array from imperfect measurements.

There are several reasons for choosing Markov fields. For one thing, there has lately been an increasing emphasis on the role of *spatial context* in image classification, in which pixels or groups of pixels are assigned symbolic labels, e.g. ground cover classes in satellite data. In these problems, contextual information is primarily local, and Markov fields provide a flexible mechanism for modeling spatial dependence. Similarly, this framework is convenient for representing other *unobserved* image attributes, especially the location of discontinuities between regions deemed homogeneous with respect to some property, such as tone, texture, or depth. Indeed these locations may be conceived as the realization of a binary process (over a "dual lattice") and thereby organized using the interactions among the "boundary variables," and between these variables and others - labels, brightness values, etc. Consider, for example, the interpolation of surfaces from noisy range measurements at sparse locations ([99]); by incorporating both range and boundary variables into the same model, it becomes possible to account explicitly for discontinuities and thereby do surface interpolation and boundary detection *at the same time*, which avoids the pitfalls (such as over-smoothing) of doing these sequentially.

Monte Carlo Methods. An unfortunate byproduct of this flexibility is that these models are usually analytically intractable. Having constrained the problem and defined the "best" solution $\hat{x}$ (according to some performance measure) there is still the problem of computing $\hat{x}$, not to mention the issue of estimating model parameters. As we have seen, these estimates are usually defined in terms of the conditional process given the data, which is another Markov field whose joint distribution is too complex for direct sampling or direct computation of global quantities such as means and modes. A partial answer is provided by algorithms designed for investigating Gibbs measures by Monte Carlo methods. This is the subject of §4; "stochastic relaxation" algorithms exploit the Markov property and analogies to statistical physics to generate sample realizations, approximate global expectations with ergodic averages, and estimate modes by "annealing." In particular, sampling and estimation are performed by the *same* computational mechanism so that "pattern synthesis = pattern analysis", as Grenander [68] puts it. Sections 4 and 5 deal with two detailed applications of this methodology, namely image restoration and boundary detection.

Parallel Developments. All approaches rely on powerful computers, due to the enormous dimensionality of the data, models, and optimization problems. Efficient use

of new, parallel hardware entails the development of correspondingly parallel (and "distributed") software, i.e. algorithms which can be executed by an array of simple and alike processing units, locally connected, independently following simple programs, yet collectively evolving towards a global equilibrium. In this regard, recent trends in image analysis mirror those in other areas, e.g. "neural networks". Finally, the predominant approach nowadays to speech recognition is based on "hidden Markov models", and there is much in common with the stochastic approach to image analysis taken in these lectures.

Organization. There are six sections and an appendix on imaging systems. The latter is intended to provide a general introduction to the physical principles of image formation and detection; the main results on blur, noise, and other sources of degradation are summarized at the beginning of §4 on image restoration. Section 2 is about discrete Markov random fields and some readers may wish to avoid the introductory material in §§2.1, 2.2 and begin with the examples and (mini-)applications in §2.3. The next section (§3) is a somewhat detailed account of stochastic relaxation and simulated annealing; all the experiments discussed and illustrated are based on the algorithms, or approximations thereof, described in §3. The next two sections are devoted to more in-depth applications: image restoration (§4), including a summary of classical methods, and boundary detection (§5), focusing on the problem of texture segmentation. Finally, §6 is given to open questions about parameter estimation, Monte Carlo algorithms, image modeling, and alternate loss functions.

Acknowledgements. I am grateful to S.Geman and J.Horowitz for comments and corrections on (incoherent) early drafts, and to C.R.Hwang and B.Gidas who brought me up-to-date on relaxation times and parameter estimation, respectively.

2 Random Fields on Graphs

We regard the entire collection of image attributes, including the brightness values and "label variables", as a network of nondeterministic states associated with the sites of a graph and interacting through its bonds. Our image models are then discrete-parameter stochastic processes; in particular, Markov random fields provide a natural setting for these spatial systems and their characterization as Gibbs distributions allows considerable flexibility for modeling and inference, as well as useful connections with statistical physics. Moreover, the notorious analytical difficulties can be partially overcome, at least from a computational viewpoint, by using the Markov property and stochastic relaxation algorithms to obtain sample realizations and to estimate distributional properties such as means and modes.

2.1 Preliminaries

The *sites* of the system are denoted by $S = \{s_1, s_2, \ldots, s_N\}$, the *state space* for the variable at site $s \in S$ by Λ_s, and the *configuration space* by Ω. Thus

$$\Omega = \prod_{s \in S} \Lambda_s, \qquad \Lambda_s \subset \mathsf{R}$$

and configurations $x \in \Omega$ are written $x = (x_s)$ or $x = (x_1, \ldots, x_N)$ for convenience, with $x_i \in \Lambda_{s_i}$, $1 \leq i \leq N$. We assume each Λ_s is countable, although we could take

Λ_s *finite* with no effect on the applications to follow. Let X_s, $s \in S$, denote the usual coordinate variables on Ω and let Π be a probability measure on Ω with $\Pi(x) > 0\ \forall x \in \Omega$. All the conditional probabilities $\Pi(X_s = x_s, s \in A \mid X_s = x_s, s \notin A)$, $A \subset S$, are then well-defined. The *local characteristics* refer to the family of univariate, conditional distributions

$$\Pi_s(\lambda \mid X_{(s)}) = \Pi(X_s = \lambda \mid X_r = x_r, r \neq s), \quad s \in S, x \in \Omega$$

where $\lambda = x_s$ and $x_{(s)} = (x_r)_{r \neq s}$.

Proposition 2.1 *The distribution of $X = (X_s)$ is determined by its local characteristics.*

Proof. We will verify that for any $x = (x_i), y = (y_i)$:

$$\frac{\Pi(x)}{\Pi(y)} = \prod_{i=1}^{N} \frac{\Pi(x_i | x_1, \ldots, x_{i-1}, y_{i+1}, \ldots, y_N)}{\Pi(y_i | x_1, \ldots, x_{i-1}, y_{i+1}, \ldots, y_N)} \tag{2.1}$$

where we have written $\Pi(x_i | x_1, \ldots, x_{i-1}, \ldots, y_{i+1}, \ldots, y_N)$ for $\Pi(X_i = x_i | X_j = x_j, 1 \leq j \leq i-1, X_j = y_j, i+1 \leq j \leq N)$ and X_i, x_i denote X_{s_i}, x_{s_i} respectively. Assuming (2.1) and that two probability measures Π, μ have the same local characteristics, we have $\frac{\Pi(x)}{\Pi(y)} = \frac{\mu(x)}{\mu(y)}$ which implies $\Pi = \mu$. To show (2.1) simply write

$$\Pi(x) = \frac{\Pi(x_N | x_1, \ldots, x_{N-1})}{\Pi(y_N | x_1, \ldots, x_{N-1})} \Pi(x_1, \ldots, x_{N-1}, y_N),$$

$$\Pi(x_1, \ldots, x_{N-1}, y_N) = \frac{\Pi(x_{N-1} | x_1, \ldots, x_{N-2}, y_N)}{\Pi(y_{N-1} | x_1, \ldots, x_{N-2}, y_N)} \Pi(x_1, \ldots, x_{N-2}, y_{N-1}, y_N)$$

and so forth. □

Note that Proposition 2.1 is false if $|S| = \infty$.

A *neighborhood system* is a collection $\mathcal{G} = (\mathcal{G}_s), s \in S$, with $\mathcal{G}_s \subset S, s \notin \mathcal{G}_s$, and $s \in \mathcal{G}_t \iff t \in \mathcal{G}_s$. The pair $(S, \mathcal{G})$ is then a *graph*; the vertices are the sites $s \in S$ and the edges are the pairs $\langle s, t \rangle$ where $s \in \mathcal{G}_t$.

A *Markov random field (MRF)* with respect to $\mathcal{G}$ is a process (X_s) as above with distribution $\Pi > 0$ such that

$$\Pi_s(x_s | x_{(s)}) = \Pi(x_s | x_r, r \in \mathcal{G}_s) \quad \forall s \in S, x \in \Omega$$

That is, $P(X_s = x_s | X_r = x_r, r \neq s) = P(X_s = x_s | X_r = x_r, r \in \mathcal{G}_s)$.

Remark 2.1 Any positive probability measure Π defines a MRF with respect to $\mathcal{G}_s = S - \{s\}$. However, the models are not computationally feasible unless the local characteristics are easily computed. Generally this is satisfied by having sparse graphs: $|\mathcal{G}_s| \ll |S|$; for example, only "local" interactions, in the sense that $\mathcal{G}_s$ consists of sites "close" to s. However, there are also practical examples of MRF models with interactions at many scales.

Example 2.1 Let $\{X_n, 0 \le n \le N\}$ be a Markov process with state space Λ, $P(X_0 = \lambda) = \nu(\lambda) > 0$, and transitions $P_n(\lambda, \delta) = P(X_{n+1} = \delta | X_n = \lambda) > 0 \; \forall \lambda, \delta \in \Lambda$. Define

$$\mathcal{G}_0 = \{1\}, \; \mathcal{G}_n = \{n-1, n+1\}, \; 1 \le n \le N-1, \; \mathcal{G}_N = \{N-1\}.$$

Then (X_n) is a MRF with respect to $\mathcal{G} = (\mathcal{G}_j)_{j=0}^N$. The local characteristics are:

$$\begin{aligned}
\Pi_0(x_0 | x_{(0)}) &= \frac{\nu(x_0) P_0(x_0, x_1)}{\sum_{\lambda \in \Lambda} \nu(\lambda) P_0(\lambda, x_1)} \\
\Pi_n(x_n | x_{(n)}) &= \frac{P_{n-1}(x_{n-1}, x_n) P_n(x_n, x_{n+1})}{\sum_{\lambda \in \Lambda} P_{n-1}(x_{n-1}, \lambda) P_n(\lambda, x_{n+1})}, \; 1 \le n \le N-1 \\
\Pi_N(x_N | x_{(N)}) &= P_{N-1}(x_{N-1}, x_N).
\end{aligned}$$

The converse, namely that the two-sided Markov property implies the one-sided Markov property, is also true. More on this later.

A *Gibbs random field* is a representation for a positive measure Π motivated by equilibrium studies in statistical physics. First we need some definitions. A *potential* is a collection $V = \{V_A : A \subset S\}$, $V_A : \Omega \to \mathbb{R}$, such that $V_\phi = 0$ and $V_A(x) = V_A(x')$ if $x_s = x'_s$ for all $s \in A$. V is *normalized* if $V_A(x) = 0$ whenever $x_t = 0$ for some $t \in A$, where we assume $0 \in \Lambda_s \; \forall_s$, although any other distinguished point would do equally well. (This condition is only imposed to insure unique representations; it has no practical importance.) The *energy* associated with V is

$$H(x) = H_V(x) = -\sum_{A \subset S} V_A(x)$$

Given a neighborhood system $\mathcal{G} = (\mathcal{G}_s)$, a *clique* is a set $C \subset S$ such that $s, t \in C, s \ne t$, imply $s \in \mathcal{G}_t$. Let $\mathbb{C}$ denote the class of cliques. A *Gibbs distribution* with respect to $\mathcal{G}$ is a measure of the form

$$\Pi(x) = Z^{-1} e^{-H(x)}, \qquad Z = \sum_x e^{-H(x)}$$

such that $Z < +\infty$ if $|\Omega| = \infty$ and V is a *$\mathcal{G}$-potential*, i.e., $V_A = 0 \forall A \notin \mathbb{C}$ and $H(x) = -\sum_{C \in \mathbb{C}} V_C(x)$.

With few exceptions, the *partition function* Z is intractable both analytically and numerically. Typically, there are parameters $\theta = (\theta_1, \ldots, \theta_J)$ in V, so that

$$Z = Z(\theta) = \sum_{x \in \Omega} e^{-H(x;\theta)}$$

The special case $H(x;\theta) = \sum_{j=1}^J \theta_j H_j(x)$ is an example of an "exponential family".

Example 2.2 1-D Ising Model. This is the Markov chain with $\Lambda = \{-1, 1\}$ and stationary transitions

$$P(x_n, x_{n+1}) = e^{\beta x_n x_{n+1}} / (e^\beta + e^{-\beta}).$$

Under the stationary initial distribution $(\frac{1}{2}, \frac{1}{2})$, one Gibbs representation has energy

$$H(x) = -\beta \sum_{i=0}^{N-1} x_i x_{i+1} \qquad (Z(\beta) = 2(e^\beta + e^{-\beta})^N)$$

The transformation $x_i \to \frac{x_i + 1}{2}$ converts Λ to $\{0, 1\}$.

Example 2.3 2-D Ising and related models.

i *Original model (Ising [88]).*

Let $S = \{(i,j): 1 \le i,j \le N\}$ and let $\mathcal{G}$ be the nearest-neighbor system: $\mathcal{G}_{i,j} = \{(i-1,j), (i+1,j), (i,j-1), (i,j+1)\} \cap S$. (The toroidal structure is simpler but unnatural for most applications.) In the classical Ising model, $\Lambda_{i,j} \equiv \{-1,+1\}$, corresponding to "spin up", "spin down", and

$$H(x) = -\frac{h}{T}\sum_s x_s - \frac{J}{T}\sum_{\langle s,t\rangle} x_s x_t,$$

where $\langle s,t\rangle$ denotes a nearest-neighbor pair, T stands for "temperature", h and J are the external magnetic field strength and coupling strength, respectively, and various other constants (e.g. materials coefficient) are incorporated into h and J. The "attractive" (resp. "repulsive") case is $J > 0$ (resp. $J < 0$).

ii *Infinite-volume Ising model:*

Take $h = 0$, $J > 0$. In contrast to the 1D case, the family of finite-dimensional distributions

$$\Pi_N(x) = e^{-H_N(x)}/Z_N, \qquad x \in \{-1,1\}^{N^2}$$

is *not* consistent. If S is replaced by $S_N = \{(i,j): -N \le i,j \le N\}$ and x is augumented by fixing all $+1's$ (resp. all $-1's$) on the boundary of S_N, the resulting family of (conditional) measures $\tilde{\Pi}_N$ converges weakly ($N \to \infty$) to an ergodic measure Π_J^+ (resp. Π_J^-) with $\Pi_J^+ = \Pi_J^-$ for $J \le J_c$ but $\Pi_J^+ \ne \Pi_J^-$ for $J > J_c$. The parameter J_c corresponds to the "critical temperature"; for values $J > J_c$, the sample configurations on the infinite lattice exhibit a high degree of regularity referred to as "long-range order." In addition, the entire family $\{\lambda\Pi_J^+ + (1-\lambda)\Pi_J^-\}$, $0 \le \lambda \le 1$, has the *same* local characteristics as Π_N (or $\tilde{\Pi}_N$).

iii *General, nearest-neighbor binary model:*

$$H(x) = -\sum_s \alpha_s x_s - \sum_{s \ne r} \beta_{sr} x_s x_r, \qquad x_s \in \{0,1\}$$

where $\beta_{rs} = \beta_{sr}$ and $\beta_{rs} = 0$ unless $|s-r| = 1$. Notice that

$$\Pi_s(\lambda | x_{(s)}) \propto \exp \lambda\{\alpha_s + \sum_{r \ne s} \beta_{sr} x_r\} \qquad (\lambda = 0,1).$$

The "auto-logistic" model is a special case with $\alpha_s \equiv \alpha$, $\beta_{sr} \equiv \beta_v$ for vertical bonds, and $\beta_{sr} \equiv \beta_h$ for horizontal bonds. The isotropic case is $\beta_v = \beta_h$; the transformation $x_s \to 2x_s - 1$ then returns us to the Ising model.

iv *Spin-glass*

This is a model for the equilibrium state of glass, in which S denotes a crystal lattice in $\mathbb{R}^2$ (or $\mathbb{R}^3$), x_s the physical state of the vertex s, and the energy is

$$H(x) = \sum_{\langle s,t \rangle} \xi_{st} x_s x_t.$$

Due to impurities in the crystal, $\{\xi_{st}\}_{\langle s,t \rangle}$ is another random field, independent of X, usually taken as Gaussian, but regarded as *fixed.* Studies involve properties of the low-energy states of H which hold a.s. (ξ).

A Representation for V. In the balance of this section, and §2.2 we have borrowed from the treatment in Griffeath [69], to which we refer the reader for further details. For $x \in \Omega, A \subset S$, set

$$x^A = (x_s^A), \qquad x_s^A = \begin{cases} x_s, & s \in A \\ 0, & s \notin A. \end{cases}$$

Proposition 2.2 *Any $\Pi > 0$ is a Gibbs distribution with respect to the "canonical" potential*

$$V_A(x) = \sum_{B \subset A} (-1)^{|A-B|} \log \Pi(x^B) \tag{2.2}$$

Moreover, for any *element $s \in A$,*

$$V_A(x) = \sum_{B \subset A} (-1)^{|A-B|} \log \Pi_s(x_s^B | x_{(s)}^B) \tag{2.3}$$

where in (2.3) s is any *element of A. The representation is unique among normalized potentials.*

The proof relies on the
Möbius Inversion Formula. Let Φ, Ψ be set functions on $\mathcal{P}(\Lambda)$, $|\Lambda| < \infty$. Then

$$\Phi(A) = \sum_{B \subset A} (-1)^{|A-B|} \Psi(B) \qquad \forall A \subset \Lambda$$

if and only if

$$\Psi(A) = \sum_{B \subset A} \Phi(B) \qquad \forall A \subset \Lambda.$$

Sketch of the Proof of Proposition 2.2
1. Π is Gibbs with respect to the potential (2.2): Define

$$\begin{cases} \Psi(A) &= \log[\Pi(x^A)/\Pi(0)] \\ \Phi(A) &= V_A(x) \end{cases}$$

where x is fixed and $0 = (0, 0, \ldots)$. Assuming (2.2) and using the inversion formula for Ψ,

$$\log \frac{\Pi(x)}{\Pi(0)} = \log \frac{\Pi(x^S)}{\Pi(0)} = \Psi(S) = \sum_{B \subset S} V_B(x)$$

Thus, $\Pi(x) = \Pi(0)\, e^{-H(x)}$, where $H(x) = -\sum_{B \subset S} V_B(x)$, and $Z = (\Pi(0))^{-1} < \infty$.

2. V is normalized: For any $s \in A$,

$$\begin{aligned} V_A(x) &= \sum_{B \subset A, s \notin B} (-1)^{|A-B|} \log \Pi(x^B) + \sum_{B \subset A, s \in B} (-1)^{|A-B|} \log \Pi(x^B) \\ &= \sum_{B \subset A-s} (-1)^{|A-B|} (\log \Pi(x^B) - \log \Pi(x^{B+s})) \end{aligned}$$

If $x_s = 0$, then $x^B = x^{B+s} \Rightarrow V_A(x) = 0$.

3. (2.2)⇔(2.3). This follows from the equation in *2.*, and the identity

$$\frac{\Pi(x^B)}{\Pi(x^{B+s})} = \frac{\Pi_s(x_s^B | x_{(s)}^B)}{\Pi_s(x_s^{B+s} | x_{(s)}^{B+s})}, \qquad s \notin B$$

4. Uniqueness: Again, one applies the Möbius Formula to the pair

$$\Phi(B) = U_B(x^A), \qquad \Psi(B) = \log \frac{\Pi(x^B)}{\Pi(0)}$$

with $\Lambda = A$, where (U_A) is another normalized potential. □

Arguments similar to those above show that if U is another potential for Π, then

$$V_A(x) = \sum_{B \subset A \subset D \subset S} (-1)^{|A-B|} U_D(x^B), \qquad A \neq \phi$$

where V is given by (2.2).

2.2 Equivalence Theorem

Theorem 2.1 *Let $\mathcal{G}$ be a neighborhood system. Then Π is a Gibbs distribution w.r.t. $\mathcal{G}$ if and only if Π is a MRF w.r.t. $\mathcal{G}$, in which case $\{V_A\}$ in (2.2) is a $\mathcal{G}$-potential.*

Note: The original version is due to Hammersley and Clifford [75] and others under some restrictions; see Kinderman and Snell [93] and the references therein. The statement and proof here are essentially due to Grimmett.

Proof. Let Π have a Gibbs representation w.r.t. $\mathcal{G}$ for some V:

$$\Pi(x) = e^{-H(x)}/Z, \qquad H(x) = -\sum_{C \in \mathbb{C}} V_C(x).$$

For $x \in \Omega, s \in S, \lambda \in \Lambda_s$ let $(\lambda, x_{(s)})$ denote the configuration obtained by replacing x_s by λ:

$$(\lambda, x_{(s)})_r = \begin{cases} x_r, & r \neq s \\ \lambda, & r = s \end{cases}$$

Then

$$\begin{aligned}\Pi_s(x_s|x_{(s)}) &= \frac{\exp -H_V(x)}{\sum_{\lambda\in\Lambda_s}\exp -H_V(\lambda, x_{(s)})} \\ &= \frac{\exp\left\{\sum_{A\in\mathbb{C},s\notin A} V_A(x)+\sum_{A\in\mathbb{C},s\in A} V_A(x)\right\}}{\sum_{\lambda\in\Lambda_s}\exp\left\{\sum_{A\in\mathbb{C},s\notin A} V_A(\lambda,x_{(s)})+\sum_{A\in\mathbb{C},s\in A} V_A(\lambda,x_{(s)})\right\}} \\ &= \frac{\exp\sum_{A\in\mathbb{C},s\in A} V_A(x)}{\sum_{\lambda\in\Lambda_s}\exp\sum_{A\in\mathbb{C},s\in A} V_A(\lambda,x_{(s)})}\end{aligned}$$

since $V_A(\lambda, x_{(s)}) = V_A(x)$ if $s \notin A$. Now $A \in \mathbb{C}$ and $s \in A$ imply that $A \subset \mathcal{G}_s + s$. Hence $\Pi_s(x_s|x_{(s)})$ depends only on x_t for $t \in \mathcal{G}_s + s$, and it follows that, in fact,

$$\Pi_s(x_s|x_{(s)}) = \Pi(x_s|x_r,\ r \in \mathcal{G}_s).$$

Now suppose Π is a MRF w.r.t. $\mathcal{G}$ and let $V = (V_A)$ be the canonical potential associated with Π as in Proposition 2.2. The proof will be completed by showing that $V_A(x) = 0$ if $A \notin \mathbb{C}$.

Choose $A \notin \mathbb{C}$. Then $\exists s, t \in A$ such that $t \notin \mathcal{G}_s + s$. Then:

$$\begin{aligned}V_A(x) &= \sum_{B\subset A}(-1)^{|A-B|}\log\Pi_s(x_s^B|x_{(s)}^B) \\ &= \sum_{B\subset A-s-t}(-1)^{|A-B|}\log\Pi_s(x_s^B|x_{(s)}^B) \\ &+ \sum_{B\subset A-s-t}(-1)^{|A-(B+s)|}\log\Pi_s(x_s^{B+s}|x_{(s)}^{B+s}) \\ &+ \sum_{B\subset A-s-t}(-1)^{|A-(B+t)|}\log\Pi_s(x_s^{B+t}|x_{(s)}^{B+t}) \\ &+ \sum_{B\subset A-s-t}(-1)^{|A-(B+s+t)|}\log\Pi_s(x_s^{B+s+t})|x_{(s)}^{B+s+t}) \\ &= \sum_{B\subset A-s-t}(-1)^{|A-B|}\log\left[\frac{\Pi_s(x_s^B|x_{(s)}^B)\Pi_s(x_s^{B+s+t}|x_{(s)}^{B+s+t})}{\Pi_s(x_s^{B+s}|x_{(s)}^{B+s})\Pi_s(x_s^{B+t}|x_{(s)}^{B+t})}\right]\end{aligned}$$

But $t \notin \mathcal{G}_s + s$ implies that $\Pi_s(x_s^B|x_{(s)}^B) = \Pi_s(x_s^{B+t}|x_{(s)}^{B+t})$ and that $\Pi_s(x_s^{B+s}|x_{(s)}^{B+s}) = \Pi_s(x_s^{B+s+t}|x_{(s)}^{B+s+t})$, and consequently that $V_A(x) = 0$. □

Remark 2.2 If V is a $\mathcal{G}$-potential, we have seen in the Equivalence Theorem that

$$\begin{aligned}\Pi_s(x_s|x_{(s)}) &= Z_s^{-1}\exp\sum_{A\in\mathbb{C},s\in A} V_A(x), \\ Z_s &= \sum_{\lambda\in\Lambda_s}\exp\left[\sum_{A\in\mathbb{C},s\in A} V_A(\lambda,x_{(s)})\right]\end{aligned}$$

For sampling and estimation *this is the important prescription*, not those in Proposition 2.2.

More generally, for $D \supset \partial A = \bigcup_{t \in A} \mathcal{G}_t \backslash A$,

$$\Pi(x_s, s \in A | x_r, r \in D) = \Pi(x_s, s \in A | x_r, r \in \partial A) = Z_A^{-1} \exp \sum_{B \cap A \neq \phi} V_B(x).$$

Example 1 (again). (Markov process on $\{0, \ldots, N\}$) If $\{X_n, 0 \leq n \leq N\}$ has the two-sided Markov property, then $\{X_n\}$ is a MRF w.r.t. $\mathcal{G}_n = \{n-1, n+1\}$, and hence has a Gibbs representation w.r.t. $\mathcal{G}$. It then follows easily that $\{X_n\}$ has the usual, one-sided Markov property.

The canonical potential for the clique $A = \{n-1, n\}$ is

$$V_A(x) = \log \left[\frac{P(X_n = x_n | X_{n-1} = x_{n-1}, X_{n+1} = 0)\, P(X_n = 0 | X_{n-1} = X_{n+1} = 0)}{P(X_n = x_n | X_{n-1} = X_{n+1} = 0)\, P(X_n = 0 | X_{n-1} = x_{n-1}, X_{n+1} = 0)} \right]$$

2.3 Examples

What follows are several examples of MRFs extracted from the literature on image modeling and classification. Apparently, Gibbs fields were introduced into image analysis by Hassner and Sklansky [76], although the treatment there is mostly expository. The discussion here is not complete; the main purpose is to illustrate the types of models actually in use. Several similar examples, however, will be developed in more detail in §§4,5.

2.3.1 Binary Fields and Channel Noise (cf. Frigessi-Piccioni [42], Marroquin [98])

This example is illuminating because the image and degradation models are sufficiently simple to permit a partial analytical treatment. The model for the (true) intensity process is the free boundary Ising model on $S \subset \mathbb{Z}^2$ with states $\{-1, +1\}$, and no external field:

$$P(X = x) = Z_\beta^{-1} \exp \left(\beta \sum_{\langle s,t \rangle} x_s x_t \right)$$

where β denotes inverse temperature, Z_β is the partition function, and $\langle s, t \rangle$ denotes a nearest neighbor pair in S. The data are generated by independently flipping the spin at each site of S with a fixed probability $0 \leq \varepsilon \leq \frac{1}{2}$. This type of degradation is common in communication theory and referred to as *memoryless binary symmetric channel noise.* Thus the observed process in $Y_s = X_s W_s$, $s \in S$, where W and X are independent processes, and (W_s) is i.i.d., $P(W_s = -1) = \varepsilon = 1 - P(W_s = 1)$.

The goal is to estimate the true signal, say x^*, from the corrupted signal $y = x^* w$. There are two parameters, β and ε, to estimate and this is the principal concern in Frigessi-Piccioni [41,42], in particular constructing estimators which are *consistent* in the large graph limit ($S \uparrow \mathbb{Z}^2$) and overcoming problems induced by phase transitions.

Marroquin [98] and Poggio, Mitter and Marroquin [100] discuss the problem of choosing an estimator for x^*. Consider the posterior distribution

$$\Pi(x|y) = P(X = x | Y = y) = Z_{\beta,y}^{-1} \exp \left\{ \beta \sum x_s x_t + \frac{1}{2} \log(\frac{1-\varepsilon}{\varepsilon}) \sum x_s y_s \right\}$$

A natural estimate, advocated in the cited references, is

$$\widehat{x}_s = \begin{cases} 1 & \text{if} \quad P(x_s = 1|Y = y) \geq \frac{1}{2} \\ -1 & \text{if} \quad P(x_s = 1|Y = y) < \frac{1}{2} \end{cases}$$

which is the Bayes estimate for the loss function

$$L(x, \widehat{x}) = \sum_{s \in S} 1_{\widehat{x}_s \neq x_s}, \qquad \widehat{x} = \widehat{x}(y)$$

i.e. the *misclassification rate.* Another possibility is the posterior mode or MAP *(maximum a posteriori)* estimator

$$\begin{aligned} \widehat{x} &= \arg\max_x P(X = x|Y = y) \\ &= \arg\min_x \left\{ -\beta \sum_{\langle s,t \rangle} x_s x_t - \frac{1}{2} \log(\frac{1-\varepsilon}{\varepsilon}) \sum x_s y_s \right\} \end{aligned}$$

In either case, it is necessary to investigate properties of the measure $\Pi(\cdot|y)$. Due to the simplicity of the noise process, $\Pi(\cdot|y)$ is another *Gibbs measure over the same graph as* $\Pi(x)$. As a result, samples and distributional properties can be obtained using stochastic relaxation (in particular the "Gibbs Sampler"-§3.2.3) and this is the approach taken in the references.

2.3.2 Texture Synthesis

Texture refers to patterns or regularity in the *local spatial distribution* of gray levels. A precise definition is elusive, and the concept is certainly scale-dependent, but it is apparent that there is a regularity in the visual appearance (at normal viewing distances) of entities such as sand, grass, wool, water and wood. A similar regularity appears in ground cover classes such as forests and wheat fields when viewed from great distances, and indeed texture is often the dominant feature in satellite, medical and other data.

It has been observed that the type of regularity mentioned above (as opposed to that in tiles or brick) could be simulated by realizations of spatial processes. Cross and Jain [30] proposed the auto-binomial MRF (cf. Besag [11]) for the synthesis of "microtextures" such as sand and water. (See also Gagalowicz and Ma [43] and the references therein.) Basically, the idea is to specify a parametric family $\Pi(\cdot, \theta)$ of random fields, estimate the parameters from one or more actual texture samples, and then compare samples from the estimated model with the real textures.

Let X_s denote the color or gray level at pixel $s \in S$, a square $N \times N$ lattice; assume there are $L + 1$ grey levels denoted $\{0, 1, \ldots, L\}$. Let $\mathcal{G}^d = \{\mathcal{G}_s^d,\ s \in S\}$ be the neighborhood system

$$\mathcal{G}_s^d = \{t \in S : 0 < |s - t|^2 \leq d\}$$

Thus $d = 1$ corresponds to the "first-order" or "nearest neighbor" system consisting of the north, south, east, and west adjacent pixels to s, $d = 2$ incorporates the four diagonal adjacencies, $d = 4$ corresponds to the twelve nearest neighbors, etc. Let $\mathbb{C}_1^d = S$, and let $\mathbb{C}_j^d$, $2 \leq j \leq m(d)$, denote the distinct (up to translation) classes of *pair cliques* for the

system $\mathcal{G}^d$; for instance, $m(1) = 3$ and $m(2) = 5$. Now fix d, put $\mathbb{C}^d = \bigcup_{j=1}^{m(d)} \mathbb{C}_j^d$, and define the potential

$$V_A(x) = \begin{cases} \log(\frac{L!}{x_s!(L-x_s)!}) + \theta_1 x_s, & A = \{s\} \\ \theta_j x_s x_t, & A = \{s,t\} \in \mathbb{C}_j,\ 2 \le j \le m \\ 0, & A \notin \mathbb{C}^d \end{cases} \tag{2.4}$$

where $\theta_1, \ldots, \theta_m \in \mathbb{R}$. To see why this is called the autobinomial model, let $\mathbf{b}(x_{(s)}) = (b_1(x_{(s)}), \ldots, b_m(x_{(s)}))$ where $b_1(x_{(s)}) = 1, b_j(x_{(s)}) = x_r + x_t$ for $\{s,r\}, \{s,t\} \in \mathbb{C}_j$, $2 \le j \le m$, with appropriate modifications near the lattice boundary. Recall that $x_{(s)}$ denotes the configuration x with x_s removed. Then (2.4) implies that the local characteristics of the MRF with potential $\{V_A\}$ are given by

$$\Pi_s(\lambda | x_{(s)}) = \left(\frac{L!}{\lambda!(\mathrm{L}-\lambda)!}\right) \tau^\lambda (1-\tau)^{L-\lambda}$$

where

$$\tau = \frac{\exp\langle \theta, \mathbf{b}(x_{(s)})\rangle}{1 + \exp\langle \theta, \mathbf{b}(x_{(s)})\rangle}$$

In other words the univariate conditional distributions are binomial (L, τ). The parameters $\theta_2, \ldots, \theta_m$ control the tendency for "bonding" in various directions, and can be adjusted to influence the visual appearance of sample realizations.

There is an interesting extension developed in Acuna [1] and motivated by the remarks of Green [64], which illustrates the use of global interactions to constrain the gray level histogram. Fix a probability vector $\mu = (\mu_0, \ldots, \mu_L)$ (the target proportions of each "color") and let $p(x) = (p_0(x), \ldots, p_L(x))$ denote the color proportions in the image: $p_j(x) = N^{-2} \sum_s 1_{x_s = j}$. Consider the augmented energy

$$H_\mu(x) = H(x) + \sigma^2 N^2 \|p(x) - \mu\|^2$$

where $H = H_V$ with V in (2.4). The local characteristics are (essentially) independent of the lattice size N for large N. Naturally, one estimates μ by the *observed* proportions, although $E_{\Pi_\mu} p_j \neq \mu_j$. Whereas H_μ has global interactions, the distribution $\Pi_s(\cdot | x_{(s)})$ depends on the values over $S - \{s\}$ only through the color counts, and efficient simulation with stochastic relaxation (see §3) is still possible. Moreover, the samples generated are less sensitive to small changes in θ than in the unconstrained version; this is important because parameter estimation is difficult (see §6).

Four textures from Acuna [1] are shown in Figure 1. These are $128x128$ images generated according to various parameter settings of the (constrained) auto-binomial model by using the Gibbs sampling algorithm (§3.2.3); the sites were visited by scanning the rows and the images represent the state of the Markov chain (see §3) after fifty full sweeps of the lattice, starting from a random coloring. We refer the reader to [1] for the specific values of the parameters θ, σ, μ.

2.3.3 Single Photon Emission Tomography (cf. S. Geman and D. McClure [54])

Emission tomography is an imaging technology for assessing organic functions such as metabolic activity and local blood flow. A pharmaceutical product is combined with a

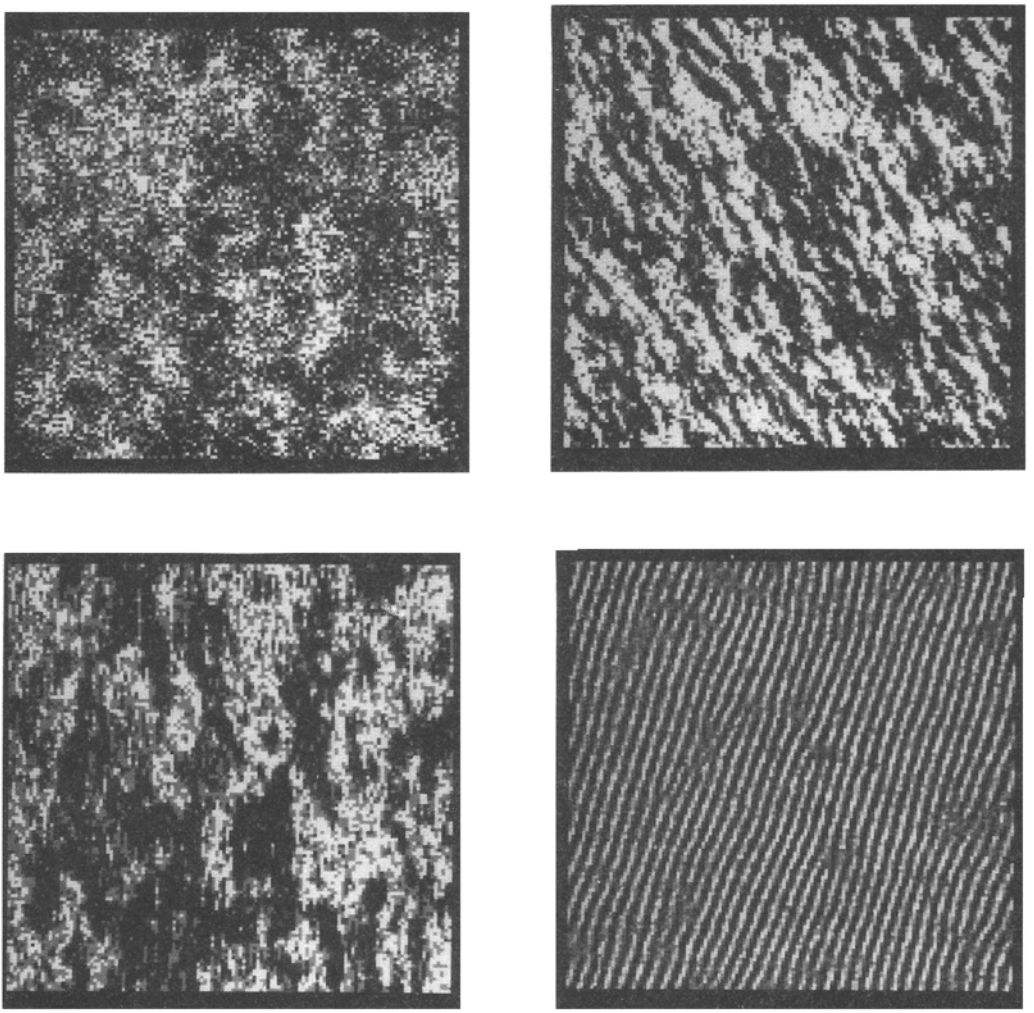


FIGURE 1

radioactive isotope and directed to a location in the body. The pharmaceutical is selected so that its concentration is proportional to the activity of interest. The objective is then to reconstruct the (internal) 2D or 3D isotope concentration based on counting the number of released photons which escape attenuation and are registered by arrays of detectors placed outside the body.

Let $X_s, s \in S$, denote the isotope density, where, for the moment, S denotes a finite planar region. The observation process is $Y_t, t \in D$, where Y_t denotes the number of photons reaching detector t. The detectors are arranged in linear arrays which can be placed at any orientation θ relative to S; consequently, it is convenient to let $D = \{(\sigma_j, \theta_k): j = 1, 2, \ldots, L,\ k = 1, \ldots, n\}$ where σ_j is the location of the j^{th} detector for the array at orientation θ_k. Assume the array is placed at n equally spaced angles for T units of time at each location. (We can imagine either one array moved about or n separate, stationary arrays.)

The degradation model $\Pi(y|x)$ is based on the physics of photon propagation and the assumption that the photons are generated by a spatially nonhomogeneous Poisson process with intensity X_s per unit time. Let $\mu(v)$ denote the attenuation coefficient at location $v \in S$; for example, if S were a planar section of the brain, then μ would assume large values near the borders of S reflecting photon absorption by the skull. Let L_t denote the line through $t \in D$ and orthogonal to the array containing t, and let $L_t(s), s \in L_t$, denote the segment of L_t between s and t. Then the probability that a photon emitted at s reaches the detector at t is

$$P \text{ (photon survives)} = \exp\left\{-\int_{L_t(s)} \mu(\ell)d\ell\right\}$$

where $d\ell$ denotes differential arc length. For idealized detectors (i.e., t a a continuous parameter), the number of photons Y_t reaching t would be a Poisson random variable with mean

$$(Rx)(t) = T\int_{L_t} x_\ell \exp\left\{-\int_{L_t(s)} \mu(\ell')d\ell'\right\} d\ell \tag{2.5}$$

This is the *attenuated Radon transform* of x. However, due to the discretization of D, and since a detector actually detects photons in small interval along the array,

$$EY_t = \int_{\theta_k - \frac{\Delta\theta}{2}}^{\theta_k + \frac{\Delta\theta}{2}} \int_{\sigma_j - \frac{\Delta\sigma}{2}}^{\sigma_j + \frac{\Delta\sigma}{2}} (Rx)(\sigma, \theta)d\sigma d\theta \tag{2.6}$$

where $\Delta\sigma$ is the spacing between detectors and the detector at θ_k counts photons in the interval $\left[\theta_k - \frac{\Delta\theta}{2}, \theta_k + \frac{\Delta\theta}{2}\right]$. In order to convert (2.6) into a *discrete* linear operator it is necessary to discretize the domain S; we assume this has been done in the usual way. Then $x_s, s \in S$ represents a piecewise constant approximation to the continuum concentration, and we can rewrite (2.6) as

$$EY_t = \sum_{s \in S} \mathcal{A}_{t,s} x_s\,; \quad EY = \mathcal{A}x$$

where $\mathcal{A}$ is a matrix of dimension $nL \times |S|$. For example, in the experiments in [54], $\mathcal{A}$ is $64^2 \times 64^2$. Finally, then, since the random variables $Y_t, t \in D$, are independent,

$$\Pi(y|x) = \prod_{t \in D} \frac{(\mathcal{A}x)_t^{y_t}}{y_t!} \exp\{-(\mathcal{A}x)_t\}. \tag{2.7}$$

Shepp and Vardi [119] applied the EM algorithm to approximate the maximum likelihood estimator (the most likely x given the counts) for the closely related problem of *positron emission tomography*. The ML estimator in our case minimizes

$$\sum_t ((\mathcal{A}x)_t - y_t \log[(\mathcal{A}x)_t])$$

which is actually a convex function of x (though not *strictly* convex), but still poses a formidable optimization problem. One motivation for the work in [54] was the tendency for this estimate to "undersmooth", i.e. produce overly irregular reconstructions.

The Bayesian approach (see also Green [65] and Dinten [32]) seeks to remedy this problem by enforcing constraints derived from the simple observation that neighboring locations in S typically have similar isotope concentrations and that sharp gradients are associated with tissue boundaries. Thus, it is important to construct a prior distribution that accounts both for local smoothness and the existence of discontinuities. For the case at hand, this will be done *implicitly*, in the sense that the locations of discontinuities will not be represented by a separate process, whereas in §4 we will see how this can be done *explicitly* with an "edge process". The prior model for x utilized in [54] to incorporate spatial information has energy

$$H(x) = \beta \sum_{\langle s,t \rangle} \phi(x_s - x_t) + \frac{\beta}{\sqrt{2}} \sum_{[s,t]} \phi(x_s - x_t)$$

where $\langle s,t \rangle$ indicates a nearest-neighbor (N,S,E,W) bond and $[s,t]$ in a diagonal bond; that is, the neighborhood system is $\mathcal{G}^2$ in which each pixel (except those near the border) has eight neighbors. The function ϕ is

$$\phi(u) = \frac{-1}{1 + (u/\delta)^2}$$

where δ is a scale parameter. The rationale for this choice is that $\phi(u) = u^2$, and other potentials for which $\lim_{u \to \infty} \phi(u) = +\infty$, will over-penalize large intensity differences and tend to suppress the formation of sharp boundaries; see §6.3. To mitigate this tendency, we want instead $\lim_{u \to \infty} \phi(u) < +\infty$, in addition to ϕ symmetric and minimized at $u = 0$. Notice that $\beta = 0$ corresponds to the i.i.d. variables $(x_s, s \in S)$ uniformly distributed over the dynamic range, whereas $\beta = +\infty$ corresponds to a degenerate law concentrated on the *constant* functions.

The posterior distribution is then

$$\Pi(x|y) = \frac{1}{Z_y} \exp\left\{ -H(x) - \sum_t ((\mathcal{A}x)_t - y_t \log((\mathcal{A}x)_t) \right\}$$

Actually, due to the complexity of the matrix $\mathcal{A}$, the graph structure for this Gibbs measure is *not* local, although relaxation algorithms of the type discussed in the following section are still feasible. Several estimates are investigated in [54]: the MAP estimator, which is the mode of $\Pi(\cdot|y)$, and the posterior mean $\widehat{x} = E(X|y)$. These correspond, respectively, to the Bayes estimates for the loss functions

$$L_{map}(x, \widehat{x}) = 1_{\{x \neq \widehat{x}\}}$$

and

$$L_{mmse}(x, \widehat{x}) = \sum_s (x_s - \widehat{x}_s)^2$$

where MMSE stands for minimum mean-squared error. We shall return to the issue of *performance measures* later in §6.4. As we have mentioned, the MAP estimator can in principle be computed with simulated annealing (see §3.3); however, some experiments in [54] are based on starting with the ML reconstruction and then using a deterministic relaxation method in which SR is "run" at temperature $T \equiv 0$, i.e. the sites of S are successively visited and the value of the field at s is replaced by the mode of the local conditional distribution $\Pi_s(x_s = \cdot | x_{(s)}, y)$ with $x_{(s)}$ fixed at the current state. The posterior mean can be approximated by exploiting the ergodicity of the Markov chain constructed in stochastic relaxation; again see §3.

Finally, there is the ever present issue of estimating the model parameters δ and β based on the actual observations y. A method of moments technique is developed in [54] due to the infeasibility of dealing with the equations arising from ML estimation with *incomplete data*; see §6.

Two experiments are shown in Figure 2. The upper two images are reconstructions based on a total of 124,136 photon counts from a cross section of a patients torso, including the lungs. The upper left image shows the maximum likelihood reconstruction; the "hot spot" (tumor) in the lung is apparent, but local structure is difficult to distinguish. The upper right shows the MMSE reconstruction with an estimated value of β.

The other two images illustrate a simulated experiment. The phantom isotope density is in the bottom left and the bottom right is a MMSE reconstruction based on simulating photon emission for a constant attenuation function with $n = L = 64$. There are sixty-four grey levels and $\delta = 12$. Again, β is estimated from the data.

2.3.4 Classification of Satellite Data

The manipulation and interpretation of remotely-sensed data comprises an enormous aspect of practical image processing. Aside from coding, storage, and enhancement, one basic issue is the identification of ground cover classes based on multispectral data, usually several visible-light "bands" and at least one in the lower frequency range. (For example, during 1972-82 the Landsat Multi-Spectral Scanner generated data in three visible bands and one infrared band.)

The principal applications involve earth resources management, for instance the monitoring of crops and forests, geologic mapping, and pollution assessment. One motivation for automated classification is the sheer amount of data: a single Landsat scene may contain 30 Mbytes of data.

"Allocation" or classification problems entail labeling the individual pixels (corresponding to small patches of ground) as belonging to one of several groups, such as "forest", "cropland", or "water", or, more specifically, individual vegetation types. Complicating factors include noisy measurements and an absence of "ground truth" (or "training samples") for proper model validation and estimation.

Statistical methods are very prevalent. However, until recently most techniques employed conventional methods, such as linear discriminant analysis, in which pixels are classified individually and independently. Starting in the 1970's and continuing with the

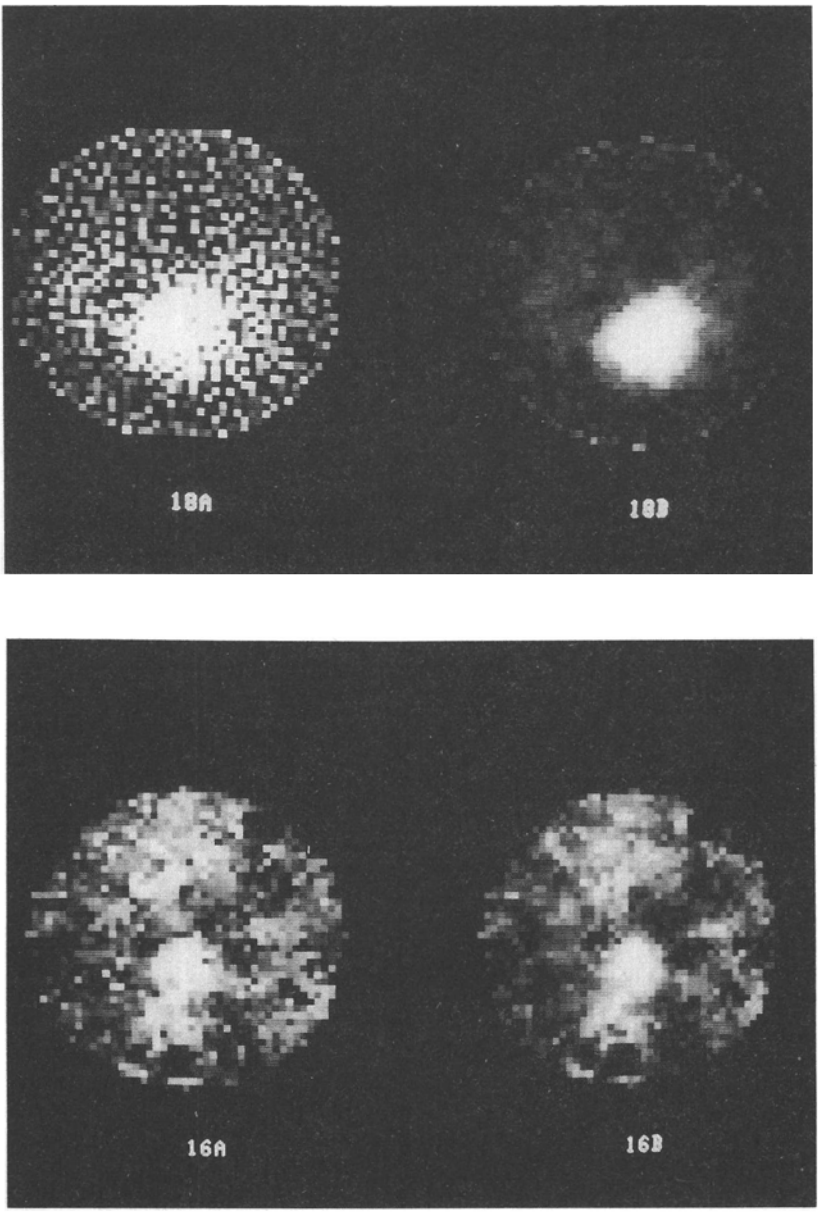


FIGURE 2

work of Swain et al [121], Besag [9], and many others, there has been an increasing emphasis on the role of *spatial context*, in which a map of labels is produced reflecting the anticipated spatial coherence; for instance, we expect crops to be grown in large homogeneous patches. We refer the reader to Ripley [116] for an account of some non-spatial methods.

The data consist of a $C \times N$ array $y = (y_1, \ldots, y_N)$, $y_n = (y_{n,1}, \cdots, y_{n,c})$, $1 \leq n \leq N$, where C is the number of bands (e.g. $C = 7$ for the current European SPOT program) and N is the number of pixels, typically very large for satellite data, say $N = (3000)^2$. (For simplicity we have relabeled the pixels $1, \ldots, N$.) Let $x^* = (x_1^*, \ldots, x_N^*)$ denote the "true" labels, where these are *unordered* and denoted simply by $\{1, 2, \ldots, M\}$. A common assumption is that y is a sample from a random vector $Y = (Y_1, \ldots, Y_N)$ and, *conditional on the true label pattern x^* the Y_i are independent, C-dimensional multivariate normal variates with mean $\mu_{x_i} = (\mu_{x_i}(1), \ldots, \mu_{x_i}(C))$ and common covariance matrix Q.* Thus:

$$\Pi(y|x) = \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}(y_i - \mu_{x_i})^T Q^{-1}(y_i - \mu_{x_i})\right\},\ y \in \mathbf{R}^C \times \mathbf{R}^N$$

From here on we put $C = 1$ for simplicity.

Following Besag [9,10], we assume x^* is a realization of a "locally dependent" Markov random field $X = (X_i)$ relative to the second order neighborhood system $\mathcal{G}^2$, and that $M = 3$ for notational convenience. Let

$$N_k = \sum_{i=1}^{N} 1_{x_i=k},\ N_{k\ell} = \sum_{\langle i,j\rangle} 1_{x_i=k,\, x_j=\ell}$$

Thus N_k is the number of pixels at "color" k and $N_{k\ell}$ is the number of neighbor pairs colored (k, ℓ). The prior label distribution is the MRF with energy

$$H(x; \alpha, \beta) = -\sum_{k=1}^{3} \alpha_k N_k(x) + \sum_{k,\ell=1, k<\ell}^{3} \beta_{k\ell} N_{k\ell}(x)$$

where $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, $\beta = (\beta_{12}, \beta_{13}, \beta_{23})$ are parameters (five of which are identifiable since $N_1 + N_2 + N_3 = N$). As usual, the partition function $Z(\alpha, \beta)$ is intractable, and hence so is the joint distribution of the variables $(N_k, N_{j\ell})$; indeed, the joint moment-generating function is

$$\begin{aligned} M(\xi, \eta) &= E\left[\exp\left(\sum \xi_k N_k + \sum \eta_{k\ell} N_{k\ell}\right)\right] \\ &= \frac{Z(\alpha, \beta)}{Z(\alpha + \xi, \beta - \eta)} \end{aligned}$$

When $\beta = 0$, the variables are independent and trinomially distributed with proportions $p_k = e^{\alpha_k} / \sum e^{\alpha_k}$; whereas when $\beta > 0$ there are strong interactions which promote label bonding.

The local characteristics are given by

$$\Pi_i(k|x_{(i)}) = \frac{\exp\left\{\alpha_k - \sum_{\ell \neq k} \beta_{k\ell}\, \nu_i(\ell)\right\}}{\sum_{k=1}^{3} \exp\left\{\alpha_k - \sum_{\ell \neq k} \beta_{k\ell}\, \nu_i(\ell)\right\}}$$

where $\nu_i(\ell)$ denotes the number of neighbors of pixel i with color ℓ. Experiments in [10] center on the special case $\beta_{k\ell} \equiv \beta$, $\alpha_k \equiv 0$, in which case

$$\Pi(x) = e^{\beta\nu(x)} / \sum_x e^{\beta\nu(x)}$$

where $\nu(x)$ counts the number of neighbor pairs which agree in color. The posterior distribution for the labels given the data has energy

$$H^P(x|y) = -\beta\nu(x) + \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mu_{x_i})^2 \tag{2.8}$$

The label estimate $\widehat{x}$ proposed in [10] (and introduced in Besag [9]) is not the MAP estimator (i.e. the minimum of (2.8)), but rather the label map obtained by *coordinate-wise minimization* in (2.8), using the neighboring values obtained in the previous steps. This algorithm is referred to as ICM for "iterated conditional modes" because one iteratively maximizes the *local* conditional distributions of the Markov random field associated with the posterior distribution:

$$\Pi_i(k|x_{(i)}, y) = Z_{i,y}^{-1} \exp\left\{\beta\nu_i(k) - \frac{1}{2\sigma^2}(y_i - \mu_k)^2\right\}$$

The ICM algorithm arrives at a *local minimum* of (2.8) and Besag discusses the advantages of exploiting only the *local* properties of the model, as well as such other issues as parameter estimation (see §6.1), block updates, and long-range order.

3 Stochastic Algorithms

3.1 Imaginary Physical Systems

It is clear from the introductory remarks and from the specific examples in the preceding chapter that estimates of image attributes and model parameters cannot be directly computed. Instead, Monte Carlo methods are used. Consider for instance the simple "smoothing" problem in which the true intensity image x^* is observed with additive noise: $y = x^* + \eta$. Constrain the solution space by regarding x^* as the realization of a MRF X; specifically, consider the MRF with Gibbs distribution

$$\Pi(x) = Z^{-1}\, e^{-H(x)}, \qquad H(x) = -\beta \sum_{\langle s,t\rangle} 1_{x_s = x_t}$$

and suppose η is white Gaussian noise and independent of X. Then the posterior distribution is

$$\Pi(x|y) = P(X = x|Y = y) = Z_y^{-1}\, e^{-H(x|y)}$$

where

$$H(x|y) = H(x) + \frac{1}{2\sigma^2}\,\|y - x\|^2$$

Two possible estimates of x^* are the mean and mode of $\Pi(x|y)$, both of which are beyond direct computation. Notice that the conditional process is a MRF over the same graph as

X; introducing correlated noise (or blur) would simply expand the neighborhood system accordingly; see §4.3. Consequently we seek Monte Carlo methods in order to simulate *global* properties of more or less arbitrary MRFs.

There is an important analogy with statistical physics. Consider a large system of basically identical, interacting components, e.g. molecules in a confined gas or atoms in a binary alloy. Let $x \in \Omega$ denote the states of the system; for example, the coordinates of x might be the molecular positions. If the system is in thermal equilibrium with its surroundings, then its behavior is determined by the *Boltzmann distribution*

$$\Pi(x) \propto \exp -\frac{1}{kT}\mathcal{H}(x)$$

k is a constant, T denotes temperature, and $\mathcal{H}$ is the Hamiltonian energy. Interest is often centered on the *ground states*

$$\Omega_{\min} = \{x: \mathcal{H}(x) = \min_x \mathcal{H}(x)\}$$

and on *ensemble averages*

$$\langle f \rangle = \int f(x)\Pi(dx)$$

Consider first these averages. In a real physical experiment, these might be directly observed (i.e. the averaging is done by nature), but this is not always possible and indirect methods are required. Ordinary Monte Carlo methods for approximating $\langle f \rangle$ involve sampling *uniformly* from Ω and averaging w.r.t. Π:

$$\langle \hat{f} \rangle = \frac{\sum_k f(x(k))\,\Pi(x(k))}{\sum_k \Pi(x(k))}$$

where $x(1), \ldots, x(k)$ are i.i.d. uniform on Ω. Notice that this estimate does not depend on the partition function, and hence can be computed explicitly. However, it will generally happen that samples $x(1), \ldots, x(k)$ have virtually zero weight under Π, unless of course k is of order $|\Omega|$, in which case the amount of computation is comparable to computing $\langle f \rangle$ directly! The observation of Metropolis et al [101] in their famous 1953 paper was that the standard procedure could essentially be *reversed* by sampling from Π and then averaging uniformly: generate a Markov chain $X(1), X(2), \ldots$ with state space Ω and asymptotic distribution Π and then estimate $\langle f \rangle$ with an ergodic average

$$\langle \hat{f} \rangle = \frac{1}{K}\sum_{k=1}^{K} f(X(k))$$

The Metropolis algorithm is a recipe for actually generating this chain and will be described below.

As for the ground states $\Omega_{\min}$, these can sometimes be reached in real physical systems by the process of chemical annealing, in which a substance is initially heated and then *slowly* cooled, allowing equilibrium to be reached at each successively lower temperature. (For example, certain crystals are obtained in this fashion.) Loosely speaking, this amounts to sampling from the Boltzmann distribution at a sequence (T_k) of decreasing temperatures. The clever idea of Kirkpatrick et al [94] and Cerny [18] (see also Pincus

[110]) was to mimic this process in order to minimize an arbitrary function H over a finite set Ω by regarding the minimizers of H as the ground states of the *imaginary* physical system with Hamiltonian energy $\mathcal{H}= H$. (Naturally, the set-up is designed for Ω formed by the product of many copies of simple sets such as $\{-1,+1\}$, and for H's which are composed of local interactions.) More specifically, the idea is to use the Metropolis algorithm to reach "steady state" at each of a sequence of temperatures T_n, $n = 1,2,3,\ldots$. For example, generate $X(1),\ldots,X(\tau_1)$ at temperature T_1, $X(\tau_1+1),\ldots,X(\tau_2)$ at temperature T_2, etc. The pair (T_n,τ_n) then constitutes an *annealing* (or *cooling) schedule.* The desired result is

$$P(X(k) \notin \Omega_{\min}) \to 0, \qquad k \to \infty.$$

We shall see how to prove this result under appropriate conditions on (T_n,τ_n) and for various transition dynamics.

3.2 Equilibrium Studies

3.2.1 Metropolis Dynamics. (cf. Hammersley and Handscomb [74])

Let Π be a probability measure on a countable set Ω with $\Pi(x) > 0\,\forall x \in \Omega$. Suppose there exists an irreducible, aperiodic Markov chain $\{X(k),\, k \geq 0\}$ with state space Ω and stationary transitions $\mathcal{P} = (P_{xy})$ such that

$$\Pi(x) = \sum_y \Pi(y) P_{yx}\,, \qquad \forall x \in \Omega. \tag{3.1}$$

Then Π is the unique invariant measure, all states are positive recurrent, and

$$\lim_{k\to\infty} P(X(k) = x \mid X(0) = y) = \Pi(x) \qquad \forall x,y \in \Omega$$

Moreover, under reasonable conditions,

$$\lim_{R\to\infty} \frac{1}{R} \sum_{k=1}^{R} f(X(k)) = \sum_x f(x)\,\Pi(x) \;\; a.s.$$

Notice that condition (3.1) can be expressed in terms of the *ratios* $\frac{\Pi(x)}{\Pi(y)}$, and so only the relative weights of Π need be given.

Construction of $\mathcal{P}$. Let Q be a *symmetric* transition matrix on $\Omega \times \Omega$, and define

$$P_{xy} = \begin{cases} Q_{xy}\,\Pi(y)/\Pi(x) & , \;\; \Pi(y) < \Pi(x) \\ Q_{xy} & , \;\; \Pi(y) \geq \Pi(x), \;\; x \neq y \\ 1 - \sum_{y\neq x} P_{xy} & , \;\; y = x \end{cases}$$

Then $\mathcal{P} = (P_{xy})$ satisfies the detailed balance or reversibility condition

$$(R) \qquad \Pi(x)P_{xy} = \Pi(y)\,P_{yx} \qquad \forall x,y$$

Notice that condition (R) implies invariance:

$$(\Pi\,\mathcal{P})_x = \sum_y \Pi(y)\,P_{yx} = \sum_y \Pi(x)\,P_{xy} = \Pi(x)$$

Alogrithm. The matrix Q is regarded as a state-generating matrix. Fix $X(0) = w$ arbitrarily. If $X(k) = x$, choose $y \in \Omega$ via Q. When $\Pi(y) \geq \Pi(x)$, we "accept" y and set $X(k+1) = y$. If $\Pi(y) < \Pi(x)$, we choose $X(k+1) = y$ with probability $\frac{\Pi(y)}{\Pi(x)}$ and we choose $X(k+1) = x$ with probability $1 - \frac{\Pi(y)}{\Pi(x)}$. Consequently,

$$P(X(k+1) = y \mid X(k) = x) = P_{xy}.$$

Moreover, Q irreducible implies $\mathcal{P}$ is irreducible (and conversely); consequently, if $\mathcal{P}$ is also aperiodic, then $X(k) \xrightarrow{D} \Pi$.

Note: If Ω is finite, then $\mathcal{P}$ is *always* aperiodic: pick $x \in \Omega$ such that $\Pi(x) \geq \Pi(y)\ \forall y$; then $P_{xx} = 1 - \sum_{y \neq x} P_{xy} = 1 - \sum_{y \neq x} Q_{xy} \frac{\Pi(y)}{\Pi(x)} > Q_{xx}$, unless $\Pi =$ const., which is of no interest.

Consider now the special case of Gibbs measures with energy H. Then the transition mechanism of $\{X(k)\}$ can be summarized as

$$P_{xy} = Q_{xy}\, e^{-[H(y)-H(x)]^+} \qquad \left(u^+ = \begin{cases} u, & u \geq 0 \\ 0, & u < 0 \end{cases}\right) \tag{3.2}$$

There are numerous open mathematical problems concerning this and related algorithms; some of these will be formulated in §6.2. One example is the *rate of convergence* to Π.

Examples of Q 1.

"Single-flip" algorithms. (These were originally designed for studying order-disorder phenomena in binary systems on subsets of $\mathbb{Z}^2$ and $\mathbb{Z}^3$.) Here $\Omega = \Lambda^S$ where $\Lambda = \{0, 1, \ldots, L\}$. Let $|S| = N$, and put

$$Q_{xy} = \begin{cases} (NL)^{-1} & , \quad |\{s : x_s \neq y_s\}| = 1 \\ 0 & , \quad \text{otherwise} \end{cases}$$

Then Q is symmetric and irreducible.

2. "Exchange" algorithms. It may be desirable to keep the overall distribution of values $\lambda \in \Lambda$ fixed, e.g. in the study of binary alloys in which there are two atomic types. One simple way is to choose at random a pair of sites $s, t \in S$ and exchange the values of the corresponding variables. Q is *not* irreducible, but the analyses above can be restricted to an equivalence class. This version of the Metropolis algorithm was employed for texture synthesis in Cross and Jain [30].

3.2.2 Generalization. (cf. Hastings [77])

Assume Ω is finite, and define $\mathcal{P} = (P_{xy})$ by

$$\begin{cases} P_{xy} &= Q_{xy}\, \alpha_{xy}, \ x \neq y \\ P_{xx} &= 1 - \sum_{y \neq x} P_{xy} \end{cases} \tag{3.3}$$

where Q is an arbitrary transition matrix and

$$\alpha_{xy} = \frac{s_{xy}}{1 + (\Pi(x)Q_{xy}/\Pi(y)Q_{yx})}$$

with s symmetric and chosen to insure that $0 \le \alpha_{xy} \le 1\ \forall x, y$.

Then the reversibility condition (R) holds, and the algorithm proceeds as before: Given $X(k) = x$, choose $y \in \Omega$ with probability Q_{xy}; set $X(k+1) = y$ with probability α_{xy} and $X(k+1) = x$ with probability $1 - \alpha_{xy}$. Then $\{X(k)\}$ has transition matrix $\mathcal{P}$.

Example 1.

$$s_{xy}^{(M)} = \begin{cases} 1 + \frac{\Pi(x)Q_{xy}}{\Pi(y)Q_{yx}} & \text{if} \quad \frac{\Pi(y)Q_{yx}}{\Pi(x)Q_{xy}} \ge 1 \\ 1 + \frac{\Pi(y)Q_{yx}}{\Pi(x)Q_{xy}} & \text{if} \quad \frac{\Pi(y)Q_{yx}}{\Pi(x)Q_{xy}} \le 1 \end{cases}$$

Then if Q is symmetric we retrieve the Metropolis algorithm.

Example 2. $s_{xy}^{(B)} \equiv 1$. Then Q symmetric implies

$$P_{xy} = Q_{xy}\left[\frac{\Pi(y)}{\Pi(x) + \Pi(y)}\right]$$

Apparently this is referred to as "Barker's method". It is equivalent to the "Gibbs Sampler" (see below) for random site-visitation and binary systems.

Remark 3.1 (cf. Erhman, Fosdick, Handscomb [39]). One may visit the sites *systematically*, changing only the designated site variable at each transition. One obtains a nonstationary Markov chain with transition matrices

$$\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_N, \qquad N = |S|,$$

which, viewed at times $t = kN, k = 0, 1, 2, \ldots$ has stationary transition matrix

$$\mathcal{P} = \prod_{j=1}^{N} \mathcal{P}_j$$

If $\Pi = \Pi\mathcal{P}_j\ \forall j$, then $\Pi = \Pi\mathcal{P}$. One must check that $\mathcal{P}$ is irreducible; otherwise Π may not be the unique equilibrium state.

3.2.3 Gibbs Sampler.

For simplicity, take the case $\Omega = \Lambda^S$, although everything works if the individual state spaces differ. Assume S is finite, say $|S| = N$.

Fix a *site-visitation schedule* $\{a_k, k = 1, 2, \ldots\}, a_k \in S$, and assume that for each $s \in S$, $a_k = s$ infinitely often. Choose $X(0)$ at random. The chain proceeds from $X(k-1)$ to $X(k)$ as follows:

$$X_s(k) = \begin{cases} X_s(k-1) & , \quad s \ne a_k \\ \xi & , \quad s = a_k \end{cases}$$

where ξ is a random variable with distribution $P(\xi = \lambda | X(j), j \le k-1) = \Pi_{a_k}(\lambda | x_{(a_k)} = X_{(a_k)}(k-1))$, i.e. the value assumed by $X_s(k)$ at $s = a_k$ is a sample from $\Pi_s(\cdot | x_r = X_r(k-1), r \ne s)$. Then $\{X(k), k \ge 0\}$ is *nonstationary* with transition matrices:

$$P_{xy}^{k} = 1_{y_{(a_k)} = x_{(a_k)}}\, \Pi_{a_k}(y_{a_k} | x_{(a_k)}) \tag{3.4}$$

Note: If Π is a MRF with respect to $\mathcal{G}$, then each transition only requires choosing a sample from the distribution on Λ given by $\Pi_s(\cdot | X_r(k-1), r \in \mathcal{G}_s)$ where s is the current site for updating.

Theorem 3.1 (Geman[2] [50])

$$\lim_{k\to\infty} P(X(k) = x \mid X(0) = y) = \Pi(x) \quad \forall x, y.$$

The proof will be given below in a slightly more general context.

Remark: Genealogy of stochastic relaxation The dynamics (3.4) is sometimes referred to as *stochastic relaxation*, although this phrase is also used for other iterative, site-replacement algorithms. The particualr algorithm here was introduced in S.Geman [51] and U.Grenander [68]; related work in somewhat different contexts appears in Glauber [59] and Ripley [114]. The first analytical treatment, including Theorem 3.1, appears in Geman[2] [50], where the algorithm is called the "Gibbs Sampler."

3.2.4 Forbidden States

It is sometimes useful (see e.g. [49], [103]) to allow states $x \in \Omega$ with $\Pi(x) = 0$ where these "forbidden states" $\Omega_F \subset \Omega$ represent estimates we wish to preclude a priori from consideration. We then consider distributions of the form

$$\Pi(x) = \Pi(x|\Omega_F^c) = 1_{x\notin\Omega_F}\, e^{-H(x)} / \sum_{x\notin\Omega_F} e^{-H(x)}$$

Example. Ω_F consists of boundary maps with undesirable properties, e.g., endings, redundant contours, etc. Or Ω_F consists of classification labels which vary too rapidly with respect to pixel units, contrary to known information about region sizes.

Still, the restriction to Ω_F^c is awkward and we can avoid it by a variation of the penalty method which generates an *irreducible* chain on Ω such that $X(k) \longrightarrow \Pi$. This subject will be treated in §3.4.

3.3 Optimization by Simulated Annealing

Part of this section is borrowed from the nice exposition in Azencott [5]. Let Ω, S, Λ be as above, with Λ *finite*. Let $H\colon \Omega \to \mathbf{R}$ be any function, let

$$\Omega_{\min} = \{x \in \Omega\colon H(x) = \min_y H(y)\},$$

and put $\Pi_T(x) = e^{-H(x)/T}/Z_T$.

Clearly, $\lim_{T\downarrow 0} \Pi_T(x) = \Pi_0(x) \doteq \delta_{\Omega_{\min}}(x) \cdot |\Omega_{\min}|^{-1}$, the uniform measure on $\Omega_{\min}$.

3.3.1 Metropolis Dynamics

Recall that, for T fixed, the Metropolis algorithm generates a Markov chain $X(0), X(1), \ldots$ with state space Ω and stationary transitions

$$P(X(k+1) = x \mid X(k) = y) = Q_{yx}\, e^{-[H(x)-H(y)]^+/T} \tag{3.5}$$

If Q is irreducible, then

$$\lim_{k\to\infty} P(X(k) = x \mid X(0) = y) = \Pi_T(x) \quad \forall x \in \Omega$$

Now fix $T_n \searrow 0$, and replace T in (3.5) by T_{k+1}.

Question 1. *Find conditions on Q and (T_k) for*

$$\lim_{k\to\infty} P(X(k) \in \Omega_{\min} \mid X(0) = y) = 1 \quad \forall y \tag{3.6}$$

Question 2. *Find conditions on Q and (T_k) for*

$$\lim_{k\to\infty} P(X(k) = x \mid X(0) = y) = \nu(x) \quad \forall x, y \tag{3.7}$$

where $\nu(x)$ is a probability measure concentrated on $\Omega_{\min}$ and $\nu(x) > 0$ on $\Omega_{\min}$.

In [50], it was shown that (3.7) holds with $\nu(x)$= uniform measure on $\Omega_{\min}$ for a related dynamics (see §3.3.2 below) if $r = \liminf T_k \log k$ is sufficiently large. Significant improvements were made, for the Metropolis dynamics, by Hajek [72], Gidas [55], Chiang and Chow [22,23] and others, e.g. characterizing r in terms of H and Q. Additional work, some involving the continuous-time (and continuous state space) analogues of $X(k)$, has been done by Hwang and Sheu [84,85,87], Chiang, Hwang and Sheu [26], Tsitsiklis [127], S.Geman and Hwang [53], Baldi [6], Mitra, Romeo, and Sangrovanni-Vincentelli [102], Jeng and Woods [91], Catoni [17], Trouvé [125], Holley and Stroock [79], Lakshmanan and Derin [96], and others.

In addition, simulated annealing has found a multitude of applications, for example in the design of integrated circuits (partitioning, placement, routing), combinatorial optimization (graph partitioning, traveling salesman problem), multiprocessor load-balancing, code compression, and statistics (e.g., cluster analysis).

Here are two representative results in the discrete case. Let $W_x = \{y \in \Omega : Q_{xy} > 0\}$, $x \in \Omega$. Since Q is symmetric, $W = \{W_x - x\}$ defines a neighborhood system on Ω. Define $x \in \Omega$ to be a *local minimum* if $H(x) \le H(y)$ for all $y \in W_x$; each $x \in \Omega_{\min}$ is a *global minimum*. Let $\Omega_{L\min}$ denote the set of local minima. We say that $x, y \in \Omega$ *communicate at height h* if either $y = x$ and $H(x) \le h$, or there is a sequence $x(1), \ldots, x(k) \in \Omega$, with $x(1) = x$ and $x(k) = y$, such that $H(x(j)) \le h$ and $x(j+1) \in W_{x(j)}$ $\forall j$.

Define the *depth* d_x of a local minimum $x \in \Omega_{L\min}$ as the smallest $D > 0$ such that there exists a $y \in \Omega$ for which x, y communicate at height $H(x) + D$ and $H(y) < H(x)$.

Theorem 3.2 (Hajek [72]) *Let*

$$D = \sup\{d_x : x \in \Omega_{L\min} - \Omega_{\min}\}$$

Then (3.6) holds if and only if

$$\sum_{k=1}^{\infty} \exp\left(-\frac{D}{T_k}\right) = +\infty.$$

Consequently, for annealing schedules of the form

$$T_k = \frac{C}{\log k}, k \ge 2$$

$P(X(k) \in \Omega_{\min}) \to 1$ if and only if $C \geq D$.

Turning to Question 2, for $x, y \in \Omega_{\min}, x \neq y$, let h_{xy} be the smallest h at which x and y communicate at height $h + H(x)$. Define

$$\overline{R} = \sup\{h_{xy}: x, y \in \Omega_{\min}, x \neq y\}$$

and

$$R = \max(\overline{R}, D)$$

Theorem 3.3 (Chiang, Chow [23]) *The condition*

$$\sum_{k=1}^{\infty} \exp\left(-\frac{R}{T_k}\right) = +\infty$$

is necessary and sufficient for (3.7).

The proofs of these results are rather technical; the latter proof involves the work of Freidlin and Wentzell on invariant measures for small diffusions. For more details, the reader is referred to the surveys of Azencott [5] and Hajek [73], the recent book of van Laarhoven and Aarts [95], and the aforementioned references.

Remark 3.2 Recently, there has been some interest in "fast annealing", i.e. schedules $T_k \searrow 0$ much faster than the logarithmic one. (For example, it has been asserted that if Q is suitably chosen than one can reduce T_k at the rate k^{-1}.) Here is a very fast annealing schedule: take Q_{xy} = uniform measure on all Ω. Then $P(X(k) \in \Omega_{\min}) \to 1$ for *any* sequence $T_k \searrow 0$! In fact, taking $T_k \equiv 0$, i.e. choosing

$$P_{xy} = Q_{xy}\, 1_{H(y)<H(x)} = \frac{1}{|\Omega|}\, 1_{H(y)<H(x)} \tag{3.8}$$

yields $P_{xx} = 1$ whenever $x \in \Omega_{\min}$, and $X(k) \to X_\infty$ with probability one, where

$$P_y(X_\infty = x) = \begin{cases} \Pi_0(x) & , \ y \notin \Omega_{\min} \\ \delta_y(x) & , \ y \in \Omega_{\min} \end{cases}$$

and Π_0 is uniform on $\Omega_{\min}$. Moreover, the expected hitting time to $\Omega_{\min}$ is finite, in fact just $|\Omega|/|\Omega_{\min}|$. Of course the algorithm in (3.8) is absurd for large-scale problems since the chain will remain in relatively high-energy states during any reasonable observation period.

3.3.2 Sampling Dynamics

The dynamics are the same as for the "Gibbs sampler" above, except that at time k we use Π_{T_k} to refresh $X_s(k-1)$ at site $s = a_k$. Thus

$$P^{\,k}_{xy} = 1_{y_{(a_k)}=x_{(a_k)}}\, \Pi_{a_k,T_k}(y_{a_k} \mid x_{(a_k)})$$

where $\Pi_{s,T}(x_s|x_{(s)})$ is the local conditional distribution relative to $\Pi_T(x) = Z_T^{-1}\, e^{-H(x)/T}$.

Theorem 3.4 (Geman[2] [50]) *Suppose* $\exists$ *an integer* $\tau \geq |S|$ *such that* $S \subseteq \{a_{k+1}, a_{k+2}, \ldots, a_{k+\tau}\}$ $\forall k = 0, 1, 2 \ldots$. *Suppose* $\liminf_{k\to\infty} T_k \log k = r$ *for some* r *sufficiently large. Then*

$$\lim_{k\to\infty} P(X(k) = x \mid X(0) = y) = \Pi_0(x) \quad \forall x, y \in \Omega.$$

Parallel Implementation. This algorithm is highly parallel: by suitably choosing $\{a_k\}$, many updates (i.e. values of the chain) can be *simultaneously* computed. The degree of parallelism is determined by the *chromatic number* of the graph associated with the MRF corresponding to Π. This is the minimum number of colors needed to label the vertices of the graph in such a way that no two vertices joined by an edge have the same color. For instance, the nearest-neighbor graph on the square lattice has chromatic number $C = 2$. Now suppose we choose $\{a_k\}$ to visit all the "black sites", then all the "red sites", etc., and imagine a processor at each site. All the black sites could be *simultaneously* refreshed, then all the red ones, etc., thereby generating $N = |S|$ values of the chain $\{X(k)\}$ in only *twice* the time required to generate a single value in a sequential realization. Obviously, the processor at site s need only be connected to those at the sites of $\mathcal{G}_s$, a modest requirement for a sparse graph.

3.4 Sampling and Annealing with Constraints

3.4.1 Problem statement

Let $H, J: \Omega \to \mathbb{R}$ where $\Omega = \Lambda^S$, $S = \{s_1, \ldots, s_N\}$, Λ is finite, and define

$$\begin{aligned} \underline{J} &= \min_{x\in\Omega} J(x), \quad \Omega(\underline{J}) = \{x \in \Omega: J(x) = \underline{J}\} \\ \underline{H} &= \min_{x\in\Omega(\underline{J})} H(x), \quad \Omega_{\min} = \{x \in \Omega(\underline{J}): H(x) = \underline{H}\} \\ \Pi(x) &= 1_{\Omega(\underline{J})}(x)\, e^{-H(x)} / \sum_{x\in\Omega(\underline{J})} e^{-H(x)} \end{aligned}$$

We consider two problems:
Problem 1: *Generate a sequence of random variables* $X(k) \to \Pi$;
Problem 2: *Solve the constrained optimization problem*

$$\text{minimize}_{\ x:J(x)=\underline{J}}\ H(x)$$

Note: Taking $J \equiv$ constant returns us to the unconstrained versions.
For each $\lambda, T > 0$, define

$$\Pi(x; T, \lambda) = \frac{\exp -\{T^{-1}(H(x) + \lambda J(x))\}}{\sum_{x\in\Omega} \exp -\{T^{-1}(H(x) + \lambda J(x))\}}$$

Simple calculations yield the following limiting behavior:

$$\begin{cases} \lim\limits_{\lambda \uparrow +\infty} & \Pi(x; 1, \lambda) = \Pi(x) \\ \lim\limits_{\substack{\lambda \uparrow +\infty \\ T \downarrow 0}} & \Pi(x; T, \lambda) = \Pi_0(x) \end{cases}$$

This asymptotic behavior suggests that we can solve problems 1 and 2 by constructing Markov chains with transitions based on $\Pi(x; T, \lambda)$.

3.4.2 Asymptotic Sampling

Consider first a more general setting. Let Π and $\{\Pi_k\}_{k=1}^{\infty}$ be any probability measures on Ω, $\Pi_k > 0$, with $\Pi_k \xrightarrow{w} \Pi$. For any $A \subset S$, $x = (x_s) \in \Omega$, let $\Pi_k(x_A|x_{(A)})$ stand for the conditional probability $\Pi_k(x_s,\ s \in A \mid x_s,\ s \notin A)$; as before, x_A denotes $(x_s)_{s\in A}$ and $x_{(A)}$ denotes $(x_s)_{s\in S\setminus A}$.

For convenience, relabel the sites $S = \{1, \ldots, N\}$. Let $\{A_k,\ k = 1, 2, \ldots\}$ be a (multiple) *site-visitation schedule*: $A_k \subset S\ \forall k$ and

$$S = \bigcup_{k=m}^{\infty} A_k \quad \forall m \geq 1.$$

Let $0 = \tau_0 < \tau_1 < \tau_2 < \ldots$ be any sequence of positive integers such that

$$S = A_{\tau_k+1} \cup A_{\tau_k+2} \cup \cdots \cup A_{\tau_{k+1}},\ k = 0, 1, 2 \ldots$$

Consider the nonstationary Markov chain on Ω with transition matrix $\mathcal{P}^k = (P^{\,k}_{xy})$ at the k^{th} step where

$$P^{\,k}_{xy} = 1_{[x_{(A_k)}=y_{(A_k)}]}\, \Pi_k(y_{A_k} \mid x_{(A_k)})$$

Thus we generate $x(k)$ from $x(k-1)$ by

$$X_s(k) = \begin{cases} X_s(k-1) & ,\ s \notin A_k \\ \xi & ,\ s \in A_k \end{cases}$$

where ξ is a sample from $\Pi_k(x_s,\ s \in A_k \mid x_s = X_s(k-1),\ s \notin A_k)$. Notice that $X(k)$ and $X(k-1)$ may differ on at most $|A_k|$ coordinates.

Finally, fix a sequence $\psi(k) \in (0,1)$, $\psi(k+1) \leq \psi(k)$, with

$$\psi(k) \leq \min_{x\in\Omega} \Pi_k(x_{A_k} \mid x_{(A_k)}) \tag{3.9}$$

and let $\|\mu\|$ denote the variation of a distribution μ on Ω.

Theorem 3.5 ([47]) *Suppose*

$$(C1) \qquad \sum_{k=1}^{\infty} \psi^N(\tau_k) = \infty$$

and

$$(C2) \qquad \sum_{k=1}^{\infty} \|\Pi_{k+1} - \Pi_k\| < +\infty$$

Then $X(k) \xrightarrow{D} \Pi$.

Remark 3.3 Suppose $|A_k| \equiv \alpha$ and S is covered (with no overlap) every m steps, i.e., $N = m\alpha$. With $\tau_k = km$, we can easily modify the proof below and replace (C1) with the condition $\sum_{k=1}^{\infty} \psi^m(km) = +\infty$, thereby getting "credit" for the block update. Clearly the "relaxation time" should decrease (i.e. the rate of convergence to equilibrium should increase) with the number of neighboring sites that are simultaneously updated; indeed, direct sampling corresponds to $A_k = S$.

Proof. In view of (C2) and standard reductions it suffices to show that $\mathbf{X} = \{X(k)\}$ is *weakly ergodic* (i.e. $\lim_{k\to\infty} \sup_{\mu,\nu} \|\mu \Pi_{j=\ell}^{k} \mathcal{P}^j - \nu \Pi_{j=\ell}^{k} \mathcal{P}^j\| = 0 \ \forall \ell \geq 1$, where μ, ν are probability vectors on Ω) and that $\Pi_k = \Pi_k \mathcal{P}^k \ \forall k$; see, e.g. Isaacson and Madsen [89, Theorem V.4.3].

The *ergodic coefficient* of a stochastic matrix Q is

$$\alpha(Q) = 1 - \frac{1}{2} \sup_{x,y} \sum_u |Q_{xu} - Q_{yu}|$$

and $\mathbf{X}$ is weakly ergodic if and only if there is a sequence $0 = \tau_0 < \tau_1 < \tau_2 < \dots$ such that

$$\sum_{k=0}^{\infty} \alpha(\mathcal{P}^{(\tau_k, \tau_{k+1})}) = +\infty \tag{3.10}$$

where $\mathcal{P}^{(m,k)} = \Pi_{j=m}^{k} \mathcal{P}^j$, $m \leq k$.

Let n_s be the last visit to site s during the epoch $(\tau_k, \tau_{k+1}]$, i.e.

$$n_s = \sup\{i \colon \tau_k < i \leq \tau_{k+1},\ s \in A_i\}, \quad 1 \leq s \leq N.$$

We can assume $n_N \leq n_{N-1} \leq \cdots \leq n_1 = \tau_{k+1}$; otherwise relabel the sites. Then

$$\begin{aligned}
\mathcal{P}^{(\tau_k, \tau_{k+1})} &= P(X(\tau_{k+1}) = y \mid X(\tau_k) = x) \\
&= P(X_s(n_s) = y_s,\ 1 \leq s \leq N \mid X(\tau_k) = x) \\
&= \prod_{s=1}^{N} P(X_s(n_s) = y_s \mid X_{s+1}(n_{s+1}) = y_{s+1}, \cdots, X_N(n_N) = y_N, X(\tau_k) = x) \\
&\geq \prod_{s=1}^{N} P(X_r(n_s) = y_r,\ r \in A_{n_s} \mid X_{s+1}(n_{s+1}) = y_{s+1}, \cdots) \\
&\geq \prod_{s=1}^{N} \psi(n_s) \\
&\geq \psi^N(\tau_{k+1}).
\end{aligned}$$

Now for any Q,

$$\alpha(Q) = \min_{x,y} \sum_u \min(Q_{xu}, Q_{yu}) \geq |\Omega| \min_{x,y} Q_{xy}$$

Hence, $\alpha(\mathcal{P}^{(\tau_k, \tau_{k+1})}) \geq |\Omega| \psi^N(\tau_{k+1})$ and (C1) implies (3.10).

Finally,

$$\begin{aligned}
(\Pi_k \mathcal{P}^k)(x) &= \sum_y \Pi_k(y) P_{yx}^k \\
&= \sum_y \Pi_k(y) 1_{x_{(A_k)} = y_{(A_k)}} \Pi_k(x_{A_k} \mid y_{(A_k)}) \\
&= \Pi_k(x_{A_k} \mid x_{(A_k)}) \sum_y \Pi_k(y) 1_{x_{(A_k)} = y_{(A_k)}} \\
&= \Pi_k(x_{A_k} \mid x_{(A_k)}) \Pi_k(x_{(A_k)}) \\
&= \Pi_k(x)
\end{aligned}$$

□

3.4.3 Back to Problems 1, 2

From here on assume

$$\Pi_k(x) = e^{-H_k(x)}/Z_k, \quad Z_k = \sum_{x\in\Omega} e^{-H_k(x)}$$

where $\{H_k\}$ satisfies the following two conditions

(I) $H_k(x) \le H_{k+1}(x) \quad \forall k \ge k_0, \ \forall x \in \Omega$ for some k_0;
(II) $\sup_k H_k(x) < +\infty$ for some $x \in \Omega$.

Since Ω is finite, (I) is equivalent to assuming that, for each x, $H_k(x)$ is eventually increasing, and (II) is equivalent to $\inf_k Z_k > 0$. Finally, assume $\tau_k = \tau\, k$ for some integer $\tau > 0$; this is harmless.

Corollary 3.1 *Let* $H_k^* = \max_x H_k(x) - \min_x H_k(x)$, *and suppose*

$$(C3) \qquad \sum_{k=1}^{\infty} \psi^N(\tau\, k) = +\infty$$

for some decreasing sequence $\psi(k)$, $\psi(k) \le e^{-H_k^*}$, $k \ge 1$. *Let* $\Pi_k \to \Pi$ *and* $\mathbf{X} = \{X(k)\}$ *be defined as above. Then*

$$\lim_{k\to\infty} P(X(k) = x \mid X(0) = y) = \Pi(x) \quad \forall x, y \in \Omega.$$

Proof. Fix $x \in \Omega$ and set $h_k = e^{-H_k(x)}$.

$$\begin{aligned} |\Pi_{k+1}(x) - \Pi_k(x)| &= \left|\frac{h_{k+1}}{Z_{k+1}} - \frac{h_k}{Z_k}\right| \\ &= (Z_{k+1}Z_k)^{-1}\ |h_{k+1}Z_k - h_kZ_{k+1}| \\ &\le (Z_{k+1}Z_k)^{-1}\{h_{k+1}|Z_{k+1} - Z_k| + Z_{k+1}|h_{k+1} - h_k|\} \\ &\le (\inf_k Z_k)^{-2}\left\{(\sup_k h_k)|Z_{k+1} - Z_k| + (\sup_k Z_k)|h_{k+1} - h_k|\right\} \end{aligned}$$

Since $0 < Z_k \searrow$ and $0 < h_k \searrow$, we then have

$$\sum_{k=1}^{\infty} |\Pi_{k+1}(x) - \Pi_k(x)| < \infty \quad \forall x$$

which implies (C2).

Finally,

$$\min_{x\in\Omega} \Pi_k(x_{A_k}|x_{(A_k)}) \ge \min_{x\in\Omega} \Pi_k(x) \ge |\Omega|^{-1}\, e^{-H_k^*}$$

and consequently (C1) holds for ψ given by (3.9). □

Note. More careful estimates, and full use of the n.a.s.c. for weak ergodicity, might yield a result similar to Hajek's.

Now take

$$H_k(x) = T_k^{-1}\{H(x) + \lambda_k\, J(x)\}$$

so that $\Pi_k(x) = \Pi(x; T_k, \lambda_k)$. Subtracting $\underline{J}$ from $J(x)$, and $\underline{H}$ from $H(x)$ (leaving Π_k unchanged), we can assume without loss of generality that $\underline{J} = 0$ and $H = 0$ on $\Omega_{\min}$.

Let $\mathbf{X} = \{X(n),\, n \geq 0\}$ correspond to (Π_k) as before, i.e. $\mathbf{X}$ has transitions

$$P_{xy}^k = 1_{x_{(A_k)}=y_{(A_k)}}\, \Pi_k(y_{A_k} \mid x_{(A_k)}),\ k \geq 1.$$

Theorem 3.6 ([47])

(a) If $T_k \equiv 1, \lambda_k \uparrow \infty$, and $\overline{\lim_k} \lambda_k (\log k)^{-1} = c$ is sufficiently small, then $X(k) \xrightarrow{D} \Pi$.

(b) If $T_k \downarrow 0, \lambda_k \uparrow \infty$, and $\overline{\lim_k} \lambda_k (T_k \log k)^{-1} = c$ is sufficiently small, then $X(k) \xrightarrow{D} \Pi_0$, the uniform measure on $\Omega_{\min}$.

Proof. (a) From the earlier results, $\Pi_k \to \Pi$. Since $H_k(x) = H(x) + \lambda_k J(x)$, we have $H_k(x) \equiv H(x)$ for $J(x) = 0$ and $H_k(x) \uparrow +\infty$ for $J(x) > 0$, satisfying (I) and (II). Let $H^* = \max_x H(x) - \min_x H(x)$ and $J^* = \max_x J(x)$. Then (C3) holds with

$$\psi(k) = \exp -(H^* + \lambda_k J^*)$$

(b) Here $H_k(x) = T_k^{-1}(H(x) + \lambda_k J(x))$. Obviously $H_k(x) \nearrow\ \forall x$ and $H_k = 0$ on $\Omega_{\min}$. Let

$$\psi(k) = \exp -(T_k^{-1}(H^* + \lambda_k J^*))$$

An easy calculation shows that (C3) is implied by $c^{-1} > NJ^*$. □

4 Image Restoration

We consider the classical image restoration problem of recovering an ideal distribution of radiant energy, $f(u), u \in \mathbf{R}^2$, from the actual recorded values $g(s), s \in \{(i,j): 1 \leq i,j \leq M\}$. We can regard g as a discrete representation of f, but degraded by blur, noise, and sampling. The elements (i,j) are referred to as "pixels"; the digital image is obtained by "quantizing" the values assumed by g to integers. The series of transformations which carry f to g is discussed in the Appendix; again, these involve distortions induced by the image formation system, the sensor, and the process of discretization. More precisely, we can regard the direct problem as the specification of the conditional distribution of the data given the true brightness values. We will assume this distribution is given. In particular, we assume the blur mechanism is known or previously estimated, which is often a reasonable assumption; for example, the blur induced by the common Vidicon camera has been extensively studied. We also assume that the noise statistics are known.

Due to the loss of information inherent in the degradation model, the "inverse problem" ranges from unstable to ill-posed. We will first review some of the standard methods such as constrained least-squares, Wiener filter, and maximum entropy which are based on a "linearization" of the problem. We will then apply the material of the preceding sections to study more general deformations and to constrain the solution space by regarding the true distribution f as the realization of a MRF and by applying Bayesian inference to estimate f.

4.1 Problem Formulation

4.1.1 Continuous-discrete model

Let $g(s)$ denote the recorded values *before quantization*; the actual data is then $y_s = [g(s)]$, where $[\,]$ denotes quantization of the range of g to the "grey levels" $k \in \{0, 1, \ldots, 2^m - 1\}$. We can simplify and summarize the discussion in the Appendix by assuming that

$$g(s) = \psi\left[\varphi(b(s)), \vec{\eta}_s\right] \tag{4.1}$$

where:

i) $b(s)$ is the intensity on the image plane at a location ξ_s (or average intensity near ξ_s) corresponding to pixel s. Assuming a linear system,

$$b(s) = \int\int f(u)K(\xi_s, u)du$$

where $K(v, u)$ is called the *point spread function* (PSF) and is the response of the image formation system at point v in the image plane to a point source of light at u in the object plane. In the space-invariant case, $K(\xi_s, u) = K(\xi_s - u)$.

ii) φ accounts for (nonlinear) sensor effects; for example, CCD (charge coupled device) cameras contain an array of small photoactive sensors; striking photons release charge carriers but the correspondence between the incident photon flux and differential electrical flux is not one-to-one. We can usually assume φ is an increasing function from $\mathbf{R}$ to $\mathbf{R}$.

iii) $\vec{\eta}$ is a collection of noise fields, including, for instance, stochastic processes corresponding to quantum and thermal fluctuations; technically, quantum noise obeys a Poisson law, but for high intensity levels the Gaussian approximation is common.

iv) ψ defines the noise mechanism, which might be signal-dependent, depending on the photon counts and other factors.

Remarks.

1. The sensor response at ξ_s may depend on $b(r)$ for points ξ_r in a neighborhood of ξ_s. However, we can imagine this effect incorporated into the point spread function K.

2. The natural domain of Fourier optics is the fully continuous formulation

$$b(v) = \int\int f(u)K(v, u)du\,, \quad v \in \mathbb{R}^2 \tag{4.2}$$

The continuous-discrete set-up we are using is more realistic; see Andrews and Hunt [4].

3. There is a body of work on "deconvolution," i.e. inverting (4.2), for various kernels K. In general the problem is not well-defined and the domain of f must be restricted to define an invertible operator. Moreover, even when an inverse exists, it generally cannot be represented as a convolution, and the problem is usually ill-conditioned in the sense

that small errors in the data may propagate to large errors in the inverse. One such example occurs in Hummel, Kimia, and Zucker [81]:

$$K(v,u) = (4\pi\, t)^{-1}\, e^{-|v-u|^2/4t}$$

where $t > 0$ represents the extent of the blur. If one regards $b(v)$ as the distribution of "heat" after t units of time, the problem is to recover the initial distribution f, i.e. to solve the heat equation *backwards in time*. The problem is regularized in [81] by restricting f to a subspace of polynomials of fixed degree.

Let us return to the general degradation model (4.1) and consider two examples.

Example 4.1 (T.V.) The general case of photoelectronic systems is reviewed in the Appendix. In this case, g corresponds to the amount of current in a scanning beam and is related to b by

$$g(s) = C(b(s))^{\gamma} + C^{1/2}(b(s))^{\gamma/2}\eta_Q(s) + \eta_t(s)$$

where b is the image formed by the camera lens, C is a constant, $\gamma > 0$ is a parameter of the detector, and η_Q, η_t represent quantum and thermal noise, respectively, which may be taken as independent, white Gaussian processes. The parameters and PSF depend on the particular imaging devices. For example, for many electro-optical devices (e.g., the Vidicon camera), the modulation transfer function (= Fourier transform of K) is of the form

$$\widehat{K}(\xi) = \exp\left[-\left(\frac{|\xi|}{\omega}\right)^{\alpha}\right]$$

where the constants α and ω depend on the particular instrument.

Example 4.2 (film) In this case, b is the irradiance distribution incident on the film and there are two ways to represent g, either as a *density image*, in which case $b > 1$ and

$$g(s) = C\, \log\, b(s) + \eta(s)$$

with η Gaussian, or as an *intensity image*, with

$$g(s) = C(b(s))^{-\gamma}\eta(s)$$

where $\gamma > 0$ and the variables $\eta(s)$ are *log-normal*. The assumptions leading to these models may be found in Andrews and Hunt [4].

4.1.2 Discrete models

For computational purposes, it is necessary to discretize f as well as g. Let us also simplify matters by ignoring the quantization and considering only one noise process. However, since we want finite lattices, and due to blur and other factors (see below), we will write $x = (x_s, s \in S)$ and $y = (y_s, s \in S')$ for the "true" image and the data, where $S = \{(i,j)\colon 1 \le i,j \le N\}$ and S' is another lattice concentric with S, but smaller, say of dimension $M \times M$ with $M \le N$. Then

$$y_s = \psi\,(\varphi((\mathcal{K}x)_s), \eta_s) \tag{4.3}$$

where $\mathcal{K}$ is a discrete representation of the PSF K:

$$(\mathcal{K}x)_s = \sum_{t \in S} \mathcal{K}(s,t)x_t\,, \qquad s \in S'.$$

Consider only the space-invariant case; then

$$(\mathcal{K}x)_s = \sum_{t \in S} \mathcal{K}(s-t)x_t \tag{4.4}$$

This is not exactly a convolution: one can imagine x_t defined on $\mathbf{Z}^2$, convolved with $\mathcal{K}$ on $\mathbf{Z}^2$, *but only observed or detected on* S' *and only reconstructed over a finite set* $S \supset S'$. If $\mathcal{K}$ has *finite* support then ideally we should take S sufficiently large to insure the summation in (4.4) includes all terms for which $\mathcal{K}(s-t) > 0$. This is a reasonable assumption in practice, since the effective blur radius rarely exceeds 10–15 pixels; for example, for Gaussian blurs, the standard deviation might range from 2–4 pixels (depending on the resolution) and hence the true PSF is well-approximated by one with a finite support. Another common assumption, but less realistic, is that (4.4) represents a circular (= toroidal) convolution; this is done to allow implementation of deconvolution filters in the Fourier domain. The method of §4.3 is based on the spatial domain and such assumptions are unnecessary.

In regard to the noise mechanism ψ, we shall assume that for each $a > 0$, the function $b \to \psi(a,b)$ is smooth and increasing; this covers the common cases $\psi(a,b) = a + b$ and $\psi(a,b) = ab$ and facilitates the computations later on.

By far the most common example of (4.3) is the *linear model*

$$y = \mathcal{K}x + \eta \tag{4.5}$$

which is often a satisfactory approximation if the noise is indeed signal-independent and if the sensor distortions are approximately linear (the so-called low-contrast assumption). There are also, however, many situations in which (4.5) is unsuitable, for instance for low signal-to-noise ratios (see [4]), and one of the attractive features of the approach taken in §4.3 is its general applicability.

4.1.3 Ill-conditioned Problems

It suffices to consider the linear model (4.5) to illustrate the problems associated with invertibility and stability. By relabeling the sites of S' and S, we can regard y, x, and η as real vectors of dimensions M^2, N^2 and M^2 respectively and $\mathcal{K}$ as an $M^2 \times N^2$ matrix, where $M^2 = |S'| \leq |S| = N^2$. It should be noticed that even if $\eta = 0$, there is at least the "noise" due to quantization: $y = [\mathcal{K}x]$ is the data vector actually recorded. Nonetheless consider first the "exact data" case $y = \mathcal{K}x$. In general, the system is then underdetermined and $\mathcal{K}^{-1}$ is obviously not well-defined. There are special cases in which one may take $N = M$, such as circular convolution or *known* values of x on the "boundary" $S \backslash S'$, but these are usually unrealistic. Moreover, even if $\mathcal{K}$ were invertible, the existence of measurement and/or quantization error renders the problem (4.5) ill-conditioned in the sense that the propagation of error from the data to the solution is not reasonably controlled. These observations can be made more precise in operator-theoretic terms, but the basic dilemma

is clear: given $\mathcal{K}$ and y, two images with blurred values very close to y can be very far apart.

Consequently, one seeks additional information to constrain the problem. We now consider several methods in which knowledge or assumptions about the intensity surfaces of the true image are exploited in the context of standard optimization theory. Finally, we shall outline another approach involving MRFs in which additional variables are introduced to account for the discontinuities between intensity surfaces.

4.2 Summary of Classical Methods

Throughout this section we assume the model (4.5) and regard $y, \eta \in \mathbf{R}^{M^2}, x \in \mathbf{R}^{N^2}$, and $\mathcal{K}$ as an $M^2 \times N^2$ matrix.

4.2.1 Least-squares or Inverse Filter

This is the simplest method: the estimate of x is

$$\widehat{x} = \arg \min_x \|y - \mathcal{K}x\|^2$$

where the notation means that $\widehat{x}$ is any minimizer of $\|y - \mathcal{K}x\|^2 = \sum(y_s - (\mathcal{K}x)_s)^2$. Thus, the least-squares estimate is any solution to the linear system

$$\mathcal{K}^t\mathcal{K}x = \mathcal{K}^t y. \tag{4.6}$$

In the usual treatment (e.g. [60]) various approximations are made to insure that $\mathcal{K}$ is a square matrix, and one also assumes that $\mathcal{K}^{-1}$ exists. The solution $\widehat{x} = \mathcal{K}^{-1}y$ is computed by approximating the block Toeplitz matrix $\mathcal{K}$ by a *block circulant* matrix, and periodically extending x and y, in order to convert matrix multiplication to circular convolution; in this way, $\widehat{x}$ may be calculated by first computing the discrete Fourier transforms (DFT) $\mathcal{F}(y)$ of y and $\mathcal{F}(\mathcal{K})$ of $\mathcal{K}$, and then applying the DFT inversion formula to the ratio $\mathcal{F}(y)/\mathcal{F}(\mathcal{K})$; see [4],[60], or [113] for definitions and details.

Generally, the problem is severely ill-conditioned. The restoration $\widehat{x} = \mathcal{K}^{-1}y = x + \mathcal{K}^{-1}\eta$ is typically dominated by the latter term because η has power concentrated in the high frequencies whereas $\mathcal{K}$, being a smoothing operator, has power concentrated in the low frequencies. Consequently, the term $\mathcal{K}^{-1}\eta$ will assume sharply oscillating values.

4.2.2 Constrained Least-squares

This is perhaps the simplest way to constrain the solution but retain the mathematical advantages of the linear/quadratic formulation. Define

$$\widehat{x} = \arg \min_{x:\|y-\mathcal{K}x\|^2=c} \|Qx\|^2 \tag{4.7}$$

where Q is a matrix and $c > 0$ is a constant chosen to reflect the noise variance σ^2. (Notice that for the true image x^*, $\|y - \mathcal{K}x^*\|^2 = \|\eta\|^2 \approx M^2\sigma^2$.)

Usually, the matrix Q is chosen to impose smoothness conditions on x, for instance,

$$\|Qx\|^2 = \sum_{s,t}(x_s - x_t)^2$$

The sum may extend over *all* pairs $s, t \in S$ or be restricted to pairs which are close together in the lattice S, thereby imposing local constraints. Higher-order smoothness conditions have also been used; for example, Hunt [82] investigated the discrete Laplacian:

$$Qx = \Delta x, \quad (\Delta x)_{ij} = x_{i+1,j} + x_{i-1,j} + x_{i,j+1} + x_{i,j-1} - 4x_{ij}$$

Experiments are reported in [82] for the case in which (η_s) consists of i.i.d. random variables uniformly distributed on $[0, \frac{1}{2}]$ (essentially quantization error) and the PSF is $\mathcal{K}(s,t) = \exp - [|s-t|^2/c]^2$; the results are impressive, and relatively easy to obtain by converting the problem to circular convolution along the lines mentioned earlier.

Note. More generally, much of the current work on "stochastic regularization" is foreshadowed by Hunt's approach to image restoration (e.g., [82,83]) which is based on random field models and Bayesian inference.

The Lagrangian formulation of (4.7) is

$$\text{minimize}_{\,x} \left[\|Qx\|^2 + \lambda \, \|y - \mathcal{K}x\|^2 \right] \tag{4.8}$$

and leads to the solution space

$$\left(\lambda^{-1} Q^t Q + \mathcal{K}^t \mathcal{K}\right) x = \mathcal{K}^t y \tag{4.9}$$

where λ must be adjusted to satisfy the contraint $\|y - \mathcal{K}x\|^2 = c$. The constrained least-squares filter is then $(\lambda^{-1} Q^t Q + \mathcal{K}^t \mathcal{K})^{-1} \mathcal{K}^t$, provided the inverse exists. Again, the most efficient implementation is in the Fourier domain after a convolution has been arranged by approximating $\lambda^{-1} Q^t Q + \mathcal{K}^t \mathcal{K}$ by a circulant matrix. However, the solutions are again often obscured by large oscillating values.

Finally, notice that the solutions to (4.9) correspond to the modes of the posterior distribution $\Pi(x|y)$ constructed from the "prior" $\Pi(x) \propto \exp(-\|Qx\|^2)$ and the conditional distribution $(\pi\lambda)^{-M^2/2} \exp(-\lambda \|y - \mathcal{K}x\|^2)$, provided we assume the noise is white Gaussian with variance $(2\lambda)^{-1}$; see, e.g. the argument in §2.3.4 or §3.1. At this point we make a digression to consider an example in some detail.

Example 4.3 Ripley [115] examines CLS for deconvolving images arising in digital astronomy. Distant galaxies are imaged with CCD cameras, and characteristically display blurring due to atmospheric distortion. Typically, the background is smooth and dark whereas the stars, which are really point sources of light, appear as bright peaks. Photons are counted by individual detectors and these counts are usually sufficiently large to warrant a Gaussian approximation, in which case the model (4.5) may be appropriate, with white Gaussian noise of variance σ_η^2. (The actual noise distribution is Poisson with mean vector $\mathcal{K}x$.) The PSF $\mathcal{K}$ is of the form

$$\mathcal{K}_{s,t} = k(|s-t|), \quad k(r) = \beta(\pi r_0^2 (1 + (\frac{r}{r_0})^2)^\beta)^{-1}$$

where $r_0 \approx 20$ and $\beta \approx 3$. (See [115] for additional comments on these assumptions.)

The prior model adopted for X is a Gaussian process expressed as a conditional autoregression (CAR):

$$\begin{aligned} E(X_s | X_t, t \neq s) &= \mu_s + \sum_t \beta_{st}(X_t - \mu_t) \quad (\beta_{tt} \equiv 0) \\ \text{var}\,(X_s | X_t, t \neq s) &= \sigma_s^2 \end{aligned}$$

with means $\mu_s \equiv 0$ and constant variance $\sigma_s^2 \equiv \sigma_x^2$. The covariance matrix is $\sigma_x^2(I - B), B = (\beta_{st})$, and B characterizes the desired neighborhood system, say $\mathcal{G}$, if we regard X as a Gibbs distribution with a continuous variables. Specifically, $B = \phi N$, where $N_{st} = 1$ if $s \in \mathcal{G}_t$ and $N_{st} = 0$ otherwise, and ϕ is a constant chosen near the common value of $|\mathcal{G}_s|^{-1}$. (The experiments reported assume a "toroidal edge connection", which is reasonable in this context due to the uniformity of the image background.) It then follows from the form of the Gaussian density and the type of degradation that

$$-2\log \Pi(x|y) = \text{ const. } + \frac{1}{\sigma_x^2} x^t(I - B)x + \frac{1}{\sigma_\eta^2} \|y - \mathcal{K}x\|^2$$

and it is easy to check that the local characteristics $\Pi_s(x_s|x_{(s)}, y)$ depend only on x_r for $r \in \mathcal{G}_s$ and those values of r for which $(\mathcal{K}^t\mathcal{K})_{sr} \neq 0$. (A more general analysis of the posterior graph follows later.)

The MAP estimator $\hat{x}$ then corresponds to (4.8) with $\lambda = \sigma_\eta^{-2}$ and $\sigma_x^2 Q^t Q = I - B$, leading to the linear system (4.9). Solutions can be obtained with the Fast Fourier Transform. Ripley also proposes an iterative method: rewrite (4.9) as $\mathcal{K}^t(y - \mathcal{K}\hat{x}) = \lambda(I - B)\hat{x}$, from which the alternative form $\hat{x} = \alpha\mathcal{K}^t y + \mathcal{A}\hat{x}$ is easily derived where $\mathcal{A} = (1-\alpha)B + \alpha(I - \mathcal{K}^t\mathcal{K}), \alpha = \sigma_x^2/(\sigma_x^2 + \sigma_\eta^2)$. The iterative step is $\hat{x}(k) = \alpha\mathcal{K}^t y + \mathcal{A}\hat{x}(k-1)$, and it can be shown that, for reasonable assumptions on $B, \mathcal{A}$ is in fact a contraction mapping. In either case, MAP deconvolution is found to be effective in this setting.

Finally, estimates of the parameters σ_η^2 and σ_x^2 are briefly considered, the latter being the more troublesome. We refer the reader to Kay [92], Thompson et al [123], and Titterington [124] for reviews of the problem of *estimating* the parameter λ, assuming one begins with the formulation (4.8), and of its effect on the reconstructions.

4.2.3 Minimum Mean Squared Error (MMSE)

Here, x and η are regarded as realizations of independent stationary processes X and $\mathcal{N}$, say both over $S' = S$ for simplicity. The MMSE estimate is then $\hat{x} = Ay$, where A minimizes the expected squared-error loss, $E\|X - AY\|^2$, over all matrices A. It is not hard to show that $A = C_x\mathcal{K}^t(\mathcal{K}C_x\mathcal{K}^t + C_\eta)$ where C_x, C_η are the covariance matrices of X and $\mathcal{N}$; see [60].

With circulant approximations, $\hat{x} \approx \mathcal{F}^{-1}(\mathcal{W}\mathcal{F}(y))$, where $\mathcal{W}$ is the *Wiener filter* and is constructed from the DFT transforms of C_x, C_η and $\mathcal{K}$. As the noise vanishes, $\mathcal{W} \to (\mathcal{F}(\mathcal{K}))^{-1}$ and we retrieve the inverse filter. Thus, for high signal-to-noise ratios, the MMSE approach is effectively (non-stochastic) least-squares estimation. On the other hand, for $\mathcal{K}$= identity (a perfect formation system), $\mathcal{W}$ reduces to the original "low-pass" filter introduced by Wiener in 1942. Ideally, it suppresses the high frequencies associated with the noise and preserves the lower ones which are assumed to comprise the image. Trouble arises when the spectral densities of the image and noise processes have overlapping supports.

4.2.4 Maximum Entropy

This is a *nonlinear* approach to the linear set-up (4.5) which has been successful in restoring certain types of images, especially those of randomly pulsed objects, starfields, etc.;

see e.g., Freiden [40] and the references in [4] and [124]. Some proponents insist that "maxent" is the *only* method of regularization that can be rationally supported; all others impose constraints derived from imprecise and hypothetical prior information, whereas in maxent the solution is derived directly from the data and "first principles".

Basically, the idea is this: The only information available about the true image x^* and noise process η^* is the constraint $y = \mathcal{K}x^* + \eta^*$, where of course y is the observed data. Moreover, let us imagine that x^* and η^* are formed by randomly distributing M_x and M_η units of energy quanta or "random grains" to N^2 pixels, subject only to the constraint $y = \mathcal{K}x^* + \eta^*$, and that these allocations for the signal and the noise are independently made. Then we seek the most likely allocations:

$$(\hat{x}, \hat{\eta}) = \arg \max_{(x,\eta): y = \mathcal{K}x + \eta} \left(\frac{M_x!}{N^{2M_x} \prod_s x_s!} \right) \left(\frac{M_\eta!}{N^{2M_\eta} \prod_s \eta_s!} \right)$$

where we assume all the quantities take integer values.

Assuming M_x, M_η and $\mathcal{K}$ are known, using Stirling's formula, and focusing only on x, the problem is often recast as

$$\hat{x} = \arg \max_{x: \|y - \mathcal{K}x\|^2 = c} \left(-\sum_s x_s \log x_s \right)$$

or, still more commonly, the term $-\sum x_s \log x_s$ is replaced by the *entropy* $-\sum_s P_s \log P_s$ where $P_s = x_s / \sum_r x_r$. Thus, $\hat{x}$ is the "most random" estimate consistent with the constraint $\|y - \mathcal{K}x\|^2 = c$; see Titterington [124] for remarks on the choice of c, as well as for the relationship between maxent and other methodologies. In particular, this approach can be interpreted as MAP estimation with the prior, (essentially) non-interacting Gibbs law $\Pi(x) \propto \exp(-\sum P_s \log P_s)$; see also Trussell [126].

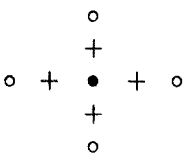
4.3 A MRF Model with Intensity Discontinuities

This section is about a model proposed in Geman[2] [50] for image restoration which accounts for both spatial coherence and the existence of discontinuities between intensity surfaces. The idea is to construct a MRF consisting of two coupled processes, one accounting for the intensity values and the other for the discontinuities or "edges". The edge process is neither part of the data nor the target of estimation. Rather, it is an auxiliary process which is coupled to the intensity process in such a manner that in the likely states of the joint probability distribution the intensity function is *locally smooth* (actually locally constant in the version here) with possibly sharp transitions and the locations of the discontinuities indicated by the edge process satisfy our expectations about the behavior of boundaries.

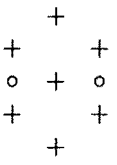
Let $S^P = \{(i,j): 1 \leq i,j \leq N\}$ and let S^E be the "dual lattice" consisting of all nearest-neighbor pairs $\langle s,t \rangle$ from S^P. The elements of S^E are referred to as "edge sites" and correspond to the location of a putative edge between the corresponding pixels. The intensity process is now denoted $X^P = \{X_s^P, s \in S^P\}$ and the edge process by $X^E = \{X_r^E, r \in S^E\}$. As usual, $X_s^P \in \Lambda = \{0, 1, \ldots, L\}$ whereas $X_r^E \in \{0,1\}$, with $X_r^E = 1$ (resp. 0) indicating the presence (resp. absence) of an edge at r.

Note: The edge sites needn't occur at the same resolution as the pixels; indeed, when the objective is boundary detection itself, rather than image restoration, it may be more effective to associate S^E with a coarse sub-lattice of S^P, as in §5.

We now consider the process $X = (X^P, X^E)$ with index set $S = S^P \cup S^E$ and configuration space $\Omega = \Lambda^{S^P} \times \{0,1\}^{S^E}$. The neighborhood system is $\mathcal{G} = \{\mathcal{G}_s, s \in S\}$; the neighborhood $\mathcal{G}_s$ of each pixel $s \in S^P$ consists of the four adjacent pixels and the four edge sites associated with s, and the neighborhood $\mathcal{G}_s$ of each edge site $s = \langle r,t \rangle \in S^E$ consists of the two pixels r,t and the six edge sites "connected" to s. (Sites near the boundary of S have fewer neighbors.) Two of these neighborhoods are illustrated below, in which the circles denote pixels and the plus signs denote edge sites.



$\mathcal{G}_s, s \in S^P$
$|\mathcal{G}_s| = 8$



$\mathcal{G}_s, s \in S^E$
$|\mathcal{G}_s| = 8$

The energy $H(x) = H(x^P, x^E)$ has two terms, reflecting expectations about interactions between intensities and edges (i.e. where edges "belong") and about boundary organization. Specifically,

$$H(x) = H^1(x^P, x^E) + H^2(x^E)$$

The component H^1 is constructed so that low energy states will have $x^E_{\langle r,t \rangle} = 1$ (resp. $= 0$) when the intensity difference $|x^P_r - x^P_t|$ is large (resp. small). In fact, we *break the bond* between pixels r,t when $x^E_{\langle r,t \rangle} = 1$. Thus we select

$$H^1(x^P, x^E) = \theta_1 \sum_{\langle r,t \rangle} \phi(x^P_r - x^P_t)(1 - x^E_{\langle r,t \rangle}) \tag{4.10}$$

where $\phi(0) = -1$, ϕ is even and non-decreasing on $[0, \infty)$, and $\theta_1 > 0$. This insures that when $x^P_r = x^P_t$, $x^E_{\langle r,t \rangle} = 0$ is a lower energy state for H^1 than $x^E_{\langle r,t \rangle} = 1$. Notice that a value of u such that $\phi(u) = 0$ represents an intensity difference for which we have no preference about the state of an edge, at least in regard to the interaction with x^P.

For small values of L we might choose

$$\phi(u) = \begin{cases} -1 & , \quad u = 0 \\ 1 & , \quad u \neq 0 \end{cases}$$

whereas for a larger dynamic range (e.g. $L \geq 15$), a typical choice is

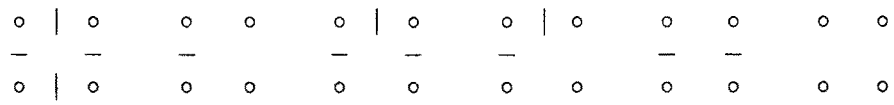
$$\phi(u) = 1 - \frac{2}{1 + (u/\delta)^2}, \; u \in [-L, L]$$

where δ is a scaling constant with $\phi(\delta) = 0$.

The organization of x^E is controlled by

$$H^2(x^E) = -\theta_2 \sum_D W_D(x^E),\ \theta_2 > 0$$

where the sum extends over all cliques D in S^E of four neighboring edge sites, the maximal cliques in the edge graph. Up to rotations of $\pi/2$, there are six possible clique states:



The slashes indicate that the edge variable at the indicated site is "on". Let $W_D = \xi_i, 1 \le i \le 6$ denote the weights assigned to these configurations. If we assume that most pixels are not at region boundaries, that edges are usually persistent, and that boundary intersections are relatively unlikely, we might choose $\xi_i \le \xi_{i+1}$, $1 \le i \le 6$. Of course these choices should depend on knowledge about the type of imagery (if available). A simple choice is $\xi_4 = \xi_5 = \xi_6 > \xi_3 > \xi_1 = \xi_2$.

Notes

i) Chalmond [19] analyses a similar model using a CAR Gaussian model for the intensities, considers estimation of the parameters based on noise-corrupted intensities, and applies the results to the restoration of NMR (nuclear magnetic resonance) images.

ii) The actual weights ξ assigned in [50] are somewhat ad hoc; a careful analysis based on polygonal approximations to continuous curves yields rather different values: Brown, Jennison, and Silverman [14].

The joint distribution of $X = (X^P, X^E)$ is then

$$\Pi(x) = \Pi(x^P, x^E) = Z^{-1} \exp -H(x^P, x^E)$$

where $Z = Z(\theta, \xi) = \sum_x e^{-H(x)}$. The degradation model is (4.3), where we take η as *white Gaussian noise* with constant mean μ, variance σ^2, and independent of X. Thus,

$$\begin{aligned}
\Pi(dy|x) &= P(Y \in dy | X = x) \\
&= P(\psi(\varphi(\mathcal{K}X^P), \eta) \in dy \mid X^P = x^P, X^E = x^E) \\
&= P(\psi(\varphi(\mathcal{K}x^P), \eta) \in dy) \\
&= \prod_s P(\psi(\varphi((\mathcal{K}x^P)_s), \eta_s) \in dy_s)
\end{aligned}$$

where each η_s is $N(\mu, \sigma^2)$. (The calculations extend easily to correlated, non-Gaussian noise.) Assume also that $\mathcal{K}$ is space-invariant and let $\mathcal{B}_s, s \in S^P$, denote the pixels which affect the blurred image $\mathcal{K}x^P$ at s, i.e. the blur support centered at s. For instance, for the simple blur

$$\mathcal{K}(s) = \begin{cases} \frac{1}{2} & , \ s = 0 \\ \\ \frac{1}{16} & , \ |s|^2 \le 2,\ s \ne 0, \end{cases}$$

$\mathcal{B}_s$ is the 3×3 square centered at s.

In view of our assumptions about $b \to \psi(a,b)$, there exists a smooth inverse $\Phi(u,v)$ which is strictly increasing in u for each v and such that

$$\eta = \Phi(y, \varphi(\mathcal{K}x^P)) = \left\{\Phi_s(y, \varphi(\mathcal{K}x^P)),\ s \in S^P\right\}$$

where in fact each Φ_s depends only on y_s and $\{x_t^P, t \in \mathcal{B}_s\}$. By space-invariance, $\mathcal{B}_{r+t} = t+\mathcal{B}_r$, where $\mathcal{B}_r \subset S^P$, $r+t \in S^P$, and $t+\mathcal{B}_r$ is intersected with S^P if necessary. Assuming $\mathcal{K}$ is symmetric, the collection $\{\mathcal{B}_s \backslash \{s\}, s \in S^P\}$ determines a neighborhood system over S^P. Let $\mathcal{B}^2$ denote the second-order system

$$\mathcal{B}_s^2 = \bigcup_{r \in \mathcal{B}_s} \mathcal{B}_r,\ s \in S^P$$

Then $\{\mathcal{B}_s^2 \backslash \{s\},\ s \in S^P\}$ is another neighborhood system. Finally, define

$$\mathcal{G}^{\mathcal{P}} = (\mathcal{G}_s^{\mathcal{P}},\ s \in S),\ \mathcal{G}_s^{\mathcal{P}} = \begin{cases} \mathcal{G}_s & ,\ s \in S^E \\ \mathcal{G}_s \cup \mathcal{B}_s^2 \backslash \{s\} & ,\ s \in S^P \end{cases}$$

Let $\Phi_1(u,v) = \frac{\partial}{\partial u}\Phi(u,v)$. Recall that $\psi(a, \Phi(b,a)) = b$, i.e. the noise can be extracted as a function of y and $\varphi(\mathcal{K}x^P)$.

Theorem 4.1 *For each y fixed, the posterior distribution*

$$\Pi(x|y) = P(X = x|Y = y),\ x \in \Omega$$

is a Gibbs distribution w.r.t. the graph $(S, \mathcal{G}^{\mathcal{P}})$ with energy

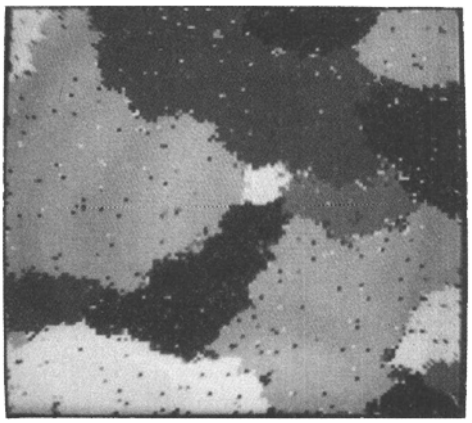
$$H^{\mathcal{P}}(x^P, x^E) = H(x^P, x^E) + \frac{1}{2\sigma^2}\left\|\vec{\mu} - \Phi(y, \varphi(\mathcal{K}x^P))\right\|^2 - \sum_{s \in S^P} \log \Phi_1(y_s, \varphi((\mathcal{K}x^P)s))$$

where $\vec{\mu} - \Phi(y, \varphi(\mathcal{K}x^P)) = \{\mu - \Phi_s\}_{s \in S^P}$.

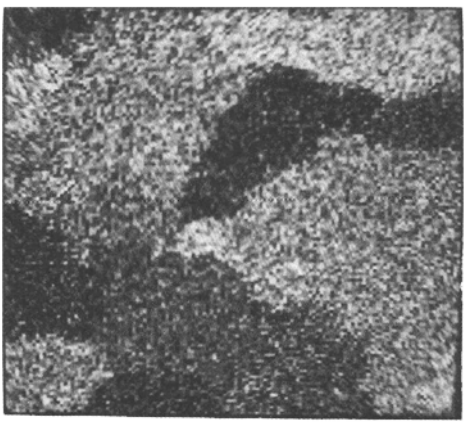
Note: The proof is essentially in [50], where the Jacobian is missing, and some technicalities about choosing regular conditional distributions are omitted.

Two experiments from [50] are shown in Figures 3 and 4. The original image in Figure 3 is a sample from a MRF over a 128x128 lattice with the eight-neighbor system; there are five grey levels and the potential function is $\phi(x_s - x_t) = -1/3$ if $x_s = x_t$, $= 1/3$ otherwise. The image was obtained by running the Gibbs sampling algorithm (§3.2.3) with cyclic site visitations; it is possible that equilibrium was not reached and would typically result in larger regions. Gaussian noise was added with $\mu = 0$ and $\sigma = 1.5$ relative to the grey levels $1, \ldots, 5$; consequently, the signal-to-noise ratio is very low. The reconstruction is based on (serial) simulated annealing (§3.3.2) with the logarithmic temperature schedule $T_k = C/log(1+k)$ where T_k is the temperature during the k'th sweep, $C = 3$, and $1 \le k \le 300$. There is no line process.

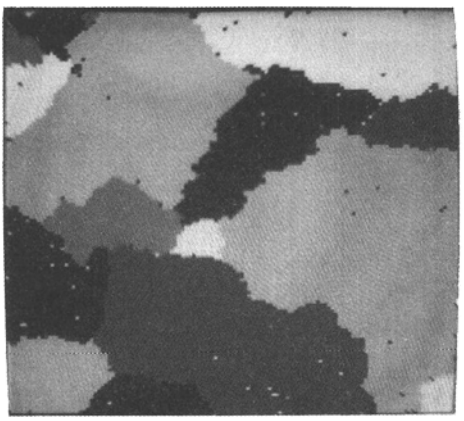
The second experiment (Figure 4) involves a hand-drawn, 64x64 image with three grey levels. The degradation is based on the model (4.3) with $\varphi(u) = u^{1/2}$, $\psi(a,b) = ab$; thus $y_s = ((\mathcal{K}x)_s)^{1/2} \cdot \eta_s$. The point spread function $\mathcal{K}$ places mass $1/2$ on the central pixel



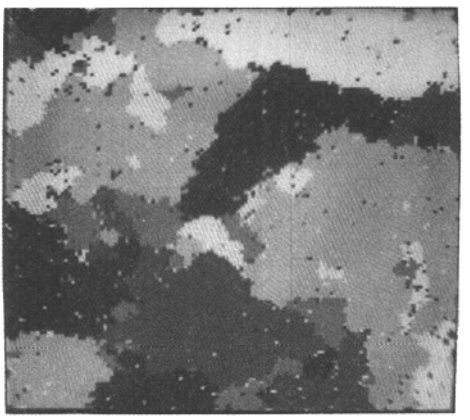
ORIGINAL



DEGRADED



RESTORATION :
300 ITERATIONS



RESTORATION :
25 ITERATIONS

FIGURE 3

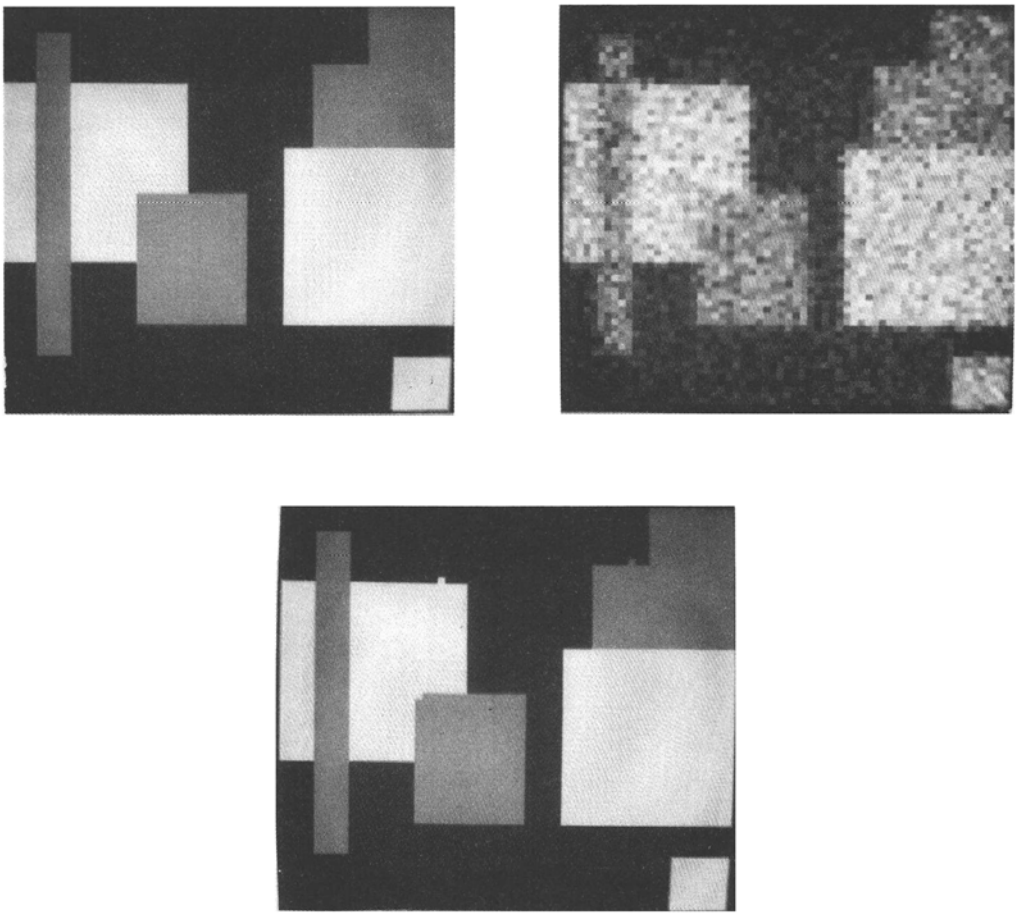


FIGURE 4

and mass 1/16 on the eight nearest neighbors, and the noise is Gaussian with $\mu = 1$ and $\sigma = 0.1$. The reconstruction is based on the same annealing algorithm as above (1000 sweeps), but here the prior model is the coupled edge/intensity process with $\phi(u) = -1$ if $u = 0, \phi(u) = 1$ if $u \neq 0$, and parameters $\theta_1 = 1, \theta_2 = 0.9$ and line clique weights $\xi_1 = 0, \xi_2 = 1, \xi_3 = \xi_4 = 2, \xi_5 = \xi_6 = 3$. The reason for favoring straight lines is obvious and illustrates the use of prior knowledge.

We refer the reader to [50] for other experiments along these lines. In addition, there are a number of papers in the references which contain other applications and innovations, including the use of Gaussian priors (Chalmond [19], Ripley [115], Goutsias and Mendel [61], Chellappa et al [21]), simultaneous parameter estimation and restoration (Lakshmanan and Derin [96], Chalmond [20], Younes [130,131]), motion data (Murray and Buxton [105]), multi-resolutions (Gidas [56]), coupled and hierarchical Markov random fields (Cohen and Cooper [27], Marroquin [98,99], Marroquin, Mitter, and Poggio [100], Jeng and Woods [88], Woods et al [128], Gamble and Poggio [44], Terzopoulos [122]), and other optimization techniques (Derin and Elliott [35], Derin et al [36], Blake [13]).

5 Boundary Detection

5.1 Physical and Digital Boundaries

The problem is to find locations in the digital image which correspond to physical discontinuities in the actual scene. These discontinuities may be sudden changes in depth (occluding boundaries), shape (creases, corners, etc.) or surface composition (texture boundaries). Complications arise because digital boundaries are often quite noisy, depending on the lighting, resolution, and other aspects of the imaging system. Consequently, it is commonplace for a sharp 3D boundary to appear in the image as a slow transition, or even sporadically disappear. In addition, pronounced intensity gradients may appear due to sensor noise, digitization, lighting, and surface irregularities. These "non-physical" edges, such as shadows, are essentially impossible to distinguish from the actual boundaries, at least without additional information from multiple sensors, temporal sequences, or specific scene knowledge, in which case boundary *classification* might be possible.

Image segmentation is a closely related and widely studied problem in both image processing and computer vision, e.g., robotics. The aim is to partition the image into disjoint classes of pixels based on local properties such as color, depth or surface orientation, or more global (even semantical) ones such as texture (e.g. vegetation types) or dichotomies such as "object-background" and "benign-malignant". Obviously each partition uniquely defines a boundary map, whereas only boundary maps which are suitably organized will yield useful segmentations.

The consensus in computer vision is that segmentation and/or boundary detection are indispensable steps towards further analyses, such as extracting geometric attributes of 3D surfaces, or interpreting images in semantical terms. In particular, algorithms for stereopsis, optical flow, and simple object recognition are often based on matching boundary segments between image pairs or between images and templates. A common image

analysis paradigm is to follow segmentation by the imposition of a relational structure among the regions; a final description results from matching these data structures against stored representations of real-world entities, such as machine parts. Some persons, including this author, have reservations about the utility of such paradigms, particularly their modular and sequential nature, and regard segmentation per se as superfluous in most real problems. On the other hand, the detection of *specific* types of boundaries can be quite useful. For example, texture (see below) is a dominant feature in remotely-sensed images and textural boundaries or texture partitions are helpful in the analysis of multi-spectral satellite data for resource classification and in geologic mapping. Other applications include automated navigation and the automated inspection of silicon wafers, in which low magnification views of memory arrays appear as highly structured textures.

We are particularly interested in the problem of texture segmentation. The given image consists of textured regions, such as a mosaic of natural textures such as sand, water and grass, or radar-imaged ice flows in water. It is well-known that humans perceive textural boundaries between regions of approximately the same average brightness because the regions themselves, although containing sharp intensity changes, are perceived as homogeneous based on properties of the *spatial distribution* of intensities. The goal then is to find these visually distinct texture regions, either by assigning categorical labels to the pixels or by detecting the boundaries directly.

The literature in these areas is immense. There are numerous algorithms for "edge detection" and image segmentation, mostly deterministic and based on heuristic arguments. Some involve the optimization of a "merit" or "goodness-of-fit" function, and some are derived from an "image model". Most general references devote sections to segmentation and edge detection; see for example Rosenfeld and Kac [118], Gonzalez and Wintz [60], and Horn [80].

The remarks in the following two subsections, §§5.2,5.3, are intended to provide some perspective; the main business is the material in §5.4, which is drawn from D.Geman, S.Geman, C.Graffigne and P.Dong [49].

5.2 Deterministic Methods

Boundaries can be regarded as well-organized subsets of edges, and "edge detection" refers to the problem of locating these individual segments, independently of the overall scene geometry, and often at the same resolution as the pixels. The construction of boundary maps is usually regarded as a second phase in which the detected edges are "cleaned" and "smoothed" and otherwise massaged into the types of structures we associate with 3D surface boundaries. Most edge detectors are based on (discrete) differential operators. One major complication is the presence of noise, and there is a fundamental trade-off between location accuracy and detection error.

Let $g(u)$ denote the given brightness distribution where, for the moment, we take $u \in \mathbf{R}^2$, and let $G_\sigma(u)$ be a smoothing kernel, e.g. $G_\sigma(u) = e^{-|u|^2/2\sigma^2}$, where σ corresponds to the degree of smoothing. Whereas the image is often already degraded by blur, the idea is to blur it *further* in order to suppress the noise. (Indeed, the standard linear approach to noise removal is just a low-pass filter.) There is one family of algorithms, "Laplacian of Gaussian," in which the edge locations $\mathbf{B}_\sigma$ are *defined* as the zero-crossings

of the Laplacian of G_σ convolved with g:

$$\begin{aligned} \mathbf{B}_\sigma &= \{u: \Delta(G_\sigma * g)(u) = 0\} \\ &= \left\{u: \left(\frac{\partial^2}{\partial u_1^2} + \frac{\partial^2}{\partial u_2^2}\right) \iint e^{-|v-u|^2/2\sigma^2} g(v)dv = 0\right\} \end{aligned}$$

Considerable attention has been devoted to the accuracy-error trade-off as a function of the "scale" σ and to the behavior of the sets $\mathbf{B}_\sigma$ as σ varies for various special cases of g. The discrete implementation involves a Gaussian "mask", the discrete Laplacian (see §4.2.2), and parameters associated with the definition of a zero-crossing. Another common approach (cf. Canny [16]) is approximated by defining the edge locations $\mathbf{B}_\sigma$ as the local maxima of $\|\nabla(G_\sigma * g)\|$ in the gradient direction $\vec{n} = \nabla(G_\sigma * g)/\|\nabla(G_\sigma * g)\|$. Again, digital implementation involves the selection of several parameters, including the blur radius σ, the scale at which a "local extremum" is defined, and, in practice, a lower bound on the magnitude of the gradient at the extrema; in fact, the set $\mathbf{B}_\sigma$ can change radically with this latter threshold but the alternative is to allow more or less *random* contributions to $\mathbf{B}_\sigma$. Indeed, the approach to boundary detection based on MRFs is partially motivated by the limitations in any sequential method originating with a *non-contextual* search for edges.

There is one other piece of work that deserves mention here. Mumford and Shah [104] develop a variational approach to a *continuum* formulation of the line process in Geman[2] [50]. Given an "irregular" image $g(u), u \in D \subset \mathbf{R}^2$, the aim is to find a pair $(f, \mathbf{B})$ where $\mathbf{B} = \{\mathbf{C}_j\}_{j=1}^J$ is a family of smooth curves, $f: D \to \mathbf{R}_+$ is smooth except on $\cup\mathbf{C}_j$, and f is "faithful" to g. The pair $(f, \mathbf{B})$ may be regarded either as a reconstruction or "cartoon" [104] of g, where the noise and texture in g are removed. This distinguished pair $(f, \mathbf{B})$ is defined as the minimum of the functional

$$\Psi_g(f, \mathbf{B}) = \alpha \iint_D (f-g)^2 + \beta \iint_{D \setminus \cup \mathbf{C}_j} \|\nabla f\|^2 + \sum_j |\mathbf{C}_j|$$

where $|\mathbf{C}_j|$ denotes the length of $\mathbf{C}_j$. The analogous discrete model is

$$H^P(x^P, x^L|y) = \alpha \sum_s (y_s - x_s^P)^2 + \beta \sum_{\langle s,t\rangle : x_{st}^L = 0} (x_s^P - x_t^P)^2 + \sum_{\langle s,t\rangle} x_{st}^L . \qquad (5.1)$$

The first term in (5.1) derives from additive white Gaussian noise ($\sigma^2 = 1/2\alpha$) and the latter two are easily seen to correspond (up to a constant) to (4.10) with $\phi(u) = \beta u^2 - 1$ and $\theta_1 = 1$. The analysis centers mostly on the one-dimensional case; the central mathematical problem is to show that the minimum *exists*.

5.3 Stochastic Image Segmentation

Some authors use the word segmentation almost interchangeably with restoration, particularly in the case of white noise and relatively few grey levels; we have already mentioned work in that area. Markov and other random field models for texture synthesis and segmentation have been fashionable for some years now. For example, in Simchony and Chellappa [120] and Derin and Cole [34] the image is modeled as a Markov random field with two layers: the "upper level" is the region process, a simple Ising-type process, and

the "lower level" is the observed intensity process, with a specified distribution (e.g., Gaussian) conditional on the region labels. These papers employ MAP and related estimators. Other methods based on random fields may be found in Derin and Elliott [35], Geman et al [48], and C.Graffigne [63].

5.4 A Markov Random Field Model for Labels

5.4.1 General case.

We are primarily interested in two specific models, one for boundary labels and the other for region labels. However, these models overlap considerably and we therefore first collect the common features. In each case, a joint distribution is specified for the labels and intensities. The degradation model is then simply a projection since we take the data to be the grey level image itself and are only interested in estimating the labels.

There are two components, one for the label-intensity interaction, which we take up first, and the other to encode constraints on the label array. Let $x^L = \{x_s^L, s \in S_\sigma^L\}$ and $x^P = \{x_s^P, s \in S^P\}$ denote the labels and the grey levels, respectively; S^P is the usual pixel lattice $\{(i,j): 1 \le i,j \le N\}$ and S_σ^L is another regular lattice, but typically more sparse, depending on the *label resolution* σ. There is a neighborhood system $\mathcal{G}$ on S_σ^L and $\langle s,t\rangle_\sigma$, $s,t \in S_\sigma^L$, denotes a neighbor pair; this system may be *non-local*, and in fact the model for region lables, or "partition model", involves a *random graph* with interactions at all scales. The x^L/x^P interaction is

$$H(x^L, x^P) = \sum_{\langle s,t\rangle_\sigma} \Psi_{s,t}(x^L)\,\Phi_{s,t}(x^P) \tag{5.2}$$

$\Psi_{s,t}(x^L)$ depends only on x_s^L and x_t^L; for example, $\Psi_{s,t}(x^L) = 1 - x_s^L x_t^L$ in the boundary model (where $x_s^L \in \{0,1\}$) and $\Psi_{s,t}(x^L) = 1_{\{x_s^L = x_t^L\}}$ in the partition model (in which case $x_s^L \in \{0,1,\ldots,P-1\}$). The other term, $\Phi_{s,t}(x^P)$, is a *measure of disparity* between the intensity values in two blocks of pixels associated with $\langle s,t\rangle_\sigma$, and will be described shortly. Large disparities ($\Phi \gg 0$) will then typically be coupled with an active boundary segment at $\langle s,t\rangle_\sigma$ ($x_s^L = x_t^L = 1$) or with dissimilar region labels ($x_s^L \ne x_t^L$) and small disparities ($\Phi \ll 0$) with an inactive boundary segment ($x_s^L x_t^L = 0$) or equal region labels ($x_s^L = x_t^L$). The effect of H is to promote placements of boundaries, or assignments of distinct labels, between regions in the image which demonstrate distinct spatial patterns as measured by Φ.

The other component in the model is a *penalty function* $J(x^L) \ge 0$ over label arrays which counts the number of "forbidden patterns" in x^L. The elements of $\{J > 0\}$ are the "forbidden states"; for instance, redundant or disconnected boundary maps or partitions with very small regions. The rationale for *hard* constraints is that our expectations about certain types of labels are quite precise and rigid. For example, most physical boundaries are smooth, persistent, and well-localized; consequently it seems reasonable to *impose* these assumptions on image boundaries, and corresponding restrictions on partition geometries.

The distribution on $x = (x^L, x^P)$ is then

$$\Pi(x^L, x^P) = 1_{\{J=0\}}(x^L)\, e^{-H(x^L,x^P)} / \sum_{x^L: J(x^L)=0} e^{-H(x^L,x^P)}. \tag{5.3}$$

As we have remarked in §3.2.4, the introduction of zero weights corresponds to "infinite energies" and the Gibbs-MRF association is altered; see Moussouris [103]. Moreover, unlike all previous applications, we assume there is no problem-specific degradation which precludes directly observing x^P. Then the data are $y = x^P$ and $\Pi(y|x)$ is singular — the point mass on $y = x^P$. To simplify the notation we shall then write $x = x^L$ and $y = x^P$. The posterior distribution at temperature T is then effectively

$$\begin{aligned} \Pi_T(x|y) &= \frac{1}{\widehat{z}_T} 1_{\{J=0\}}(x) \exp\{-H(x,y)/T\} \\ \widehat{z}_T &= \sum_{x:J(x)=0} \exp\{-H(x,y)/T\} \end{aligned} \tag{5.4}$$

The kinds of estimates we are interested in are

$$\widehat{x}_{map} = \arg \min_{x:J(x)=0} H(x,y)$$

and

$$\widehat{x}_{samp} = \text{sample from } \Pi_T(\cdot|y) \text{ for } T \text{ "small"}$$

As described in §3, these estimates are limits of Markov chains with transition functions constructed from the multivariate local characteristics of the parametric Gibbs family

$$\Pi_{T,\lambda}(x|y) = \frac{1}{\widehat{z}_{T,\lambda}} e^{-\frac{1}{T}\{H(x,y)+\lambda J(x)\}} \tag{5.5}$$

for appropriate sequences $T_k \searrow 0$ and $\lambda_k \nearrow \infty$. In other words, the (infinite) energy barriers due to the constraints $\{J = 0\}$ are introduced gradually during the relaxation process.

The critical choice is Φ. Each $\langle s,t\rangle_\sigma$ is associated with a pair of blocks of pixels, say D_1, D_2; in the partition model these are centered at s and t and may be widely separated, whereas in the boundary model the blocks are adjacent and "straddle" the boundary segment $\langle s,t\rangle_\sigma$. Our main concern is texture segmentation, in which case $\sigma > 1$ and the size of the blocks must be sufficiently large to characterize the texture, at least in comparison with the other textures and with respect to the particular spatial statistics. (The actual block size in [49] is $|D| = 441 = 21 \times 21$ for both models.) There are basically two cases: if disparity is gauged by the raw grey levels (the "first-order" case), then $\Phi_{s,t}(y)$ is constructed from the Kolmogorov-Smirnov distance

$$\nu(y(D_1), y(D_2)) = \max_{-\infty<u<+\infty} |\widehat{F}_1(u) - \widehat{F}_2(u)| \tag{5.6}$$

where

$$\widehat{F}_i(u) = \frac{1}{|D_i|} \text{ card } \{r \in D_i : y_r \le u\}$$

and $y(D_i)$ is the image data over D_i. As we shall see, the labeling $\widehat{x}$ will inherit desirable invariance properties from the Kolmogorov-Smirnov statistic. If $y(D_1)$ and $y(D_2)$ represented i.i.d. samples from two continuous distributions, then under the hypotheses that these distributions are identical, the distribution of $\nu(y(D_1), y(D_2))$ would actually

be independent of this common distribution. For the cases at hand, these assumptions are generally violated. However, the formal theory is primarily motivational; the distance ν is an effective "measure of homogeneity" with the type of invariance (to illumination changes) we are looking for.

Many visually distinct textures have nearly identical first-order statistics. In particular, we perceive textural boundaries between regions of approximately the same average brightness and variation. Thus, discrimination must rely on higher-order spatial statistics. We accomplish this by comparing the first-order statistics (i.e. the histograms) of various *data transformations*, say $\Gamma_1, \Gamma_2, \ldots, \Gamma_m$. Formally, each $\Gamma_i: y \to y^{(i)}$, where $y_s^{(i)}$ depends on *both* y_s *and* the grey levels in a "window" centered at $s \in S^P$. For example, $y_s^{(i)}$ might be the mean, range, or variance in this window, or a "directional residual" of the form

$$y_s^{(i)} = \left| y_s - \sum \alpha_j y_{s_j} \right| \tag{5.7}$$

where $\sum \alpha_j = 1$ and $\{s_j\}$ are pixels nearby to pixel s, perhaps surrounding s, or in the same row, column, or diagonal. These transforms are decidedly multivariate, depending on the marginal distributions of the original data y of at least dimension three. In contrast, most approaches to texture segmentation are based solely on the one- and two-dimensional marginals, the grey level histogram and the so-called "co-occurrence matrices", which are simply the empirical distributions of order two. We refer the reader to [49] for a more complete discussion except to emphasize that we make no claims to finding the "right" family of transformations for converting texture differences into tone differences. Indeed, our primary purpose is to provide a coherent mechanism for integrating the multiple, even redundant, "cues" determined by the Γ_i. The measures $\Phi_{s,t}$ will then be constructed from the distance

$$\max_{1 \le i \le m} \left[c_i^{-1} \, \nu(y^{(i)}(D_1), \, y^{(i)}(D_2)) \right] \tag{5.8}$$

where $y^{(i)} = \Gamma_i(y)$, $1 \le i \le m$, and $y^{(i)}(D_j) = \{y_s^{(i)}, s \in D_j\}$, $j = 1, 2$. The idea is that the blocks D_1, D_2 are disparate when at least one transform converts the textures in D_1, D_2 into spatial distributions with different first-order statistics. The thresholds $c_1, \ldots, c_m$ are chosen to limit the percentage of "false alarms": cases in which two blocks in the same texture are found to be "different". Finally, it follows from the aforementioned invariance, that all the labelings derived are invariant to at least *linear* changes in grey levels, i.e. $\hat{x}(y) = \hat{x}(ay + b)$.

5.4.2 Boundaries

The boundary process is $X = \{X_s, s \in S_\sigma^L\}$, $X_s \in \{0, 1\}$, where $S_\sigma^L \subset S^L$ is the sub-lattice

$$S_\sigma^L = \left\{ (i\sigma + 1, j\sigma + 1) : 1 \le i, j \le \frac{N-2}{\sigma} \right\} \tag{5.9}$$

and S^L is a regular lattice of dimension $(N-1) \times (N-1)$ and interspersed among the pixels. The graph structure in S_σ^L is nearest-neighbor, and we identify $\langle s, t \rangle_\sigma$ with the elementary boundary segment consisting of the horizontal or vertical string of $\sigma + 1$ sites in S^L, including s, t and the $\sigma - 1$ sites "in between". The interaction energy is

$$H(x, y) = \sum_{\langle s,t \rangle_\sigma} (1 - x_s x_t) \phi(\Delta_{s,t}(y)) \tag{5.10}$$

where $\Delta_{s,t}$ is constructed from the Kolmogorov-Smirnov distance and ϕ is chosen so that large values of $\Delta_{s,t}$ are coupled to $x_s x_t = 1$ and small values to $x_s x_t = 0$. When $\sigma = 1$, the products $x_s x_t$ correspond directly to the edge variables $x^E_{s,t}$ in §4.3. Since $y = x^P$ is *fixed* here, the term $1 - x_s x_t$ can be replaced with $-x_s x_t$ with no change.

A little reflection shows that $\phi(r), 0 \le r < \infty$, should be increasing with $\phi(0) < 0 < \phi(+\infty)$. The intercept $\beta = \phi^{-1}(0)$ is critical; values of Δ above (resp. below) β will promote (resp. inhibit) boundary formation. The influence of β is somewhat diminished by arranging $\phi'(\beta) = 0$. We have chosen

$$\phi(r) = \begin{cases} -(\frac{r-\beta}{\beta})^2 & , \quad 0 \le r \le \beta \\ (\frac{r-\beta}{\alpha-\beta})^2 & , \quad r > \beta \end{cases}$$

where $\alpha \approx \max \Delta_{s,t}$. In this way, the maximum "penalty" ($\phi(0) = -1$) and "reward" ($\phi(1) = 1$) are equalized.

The choice of $\Delta_{s,t}$ is problem-dependent. Consider first the case of detecting changes in brightness itself; thus, depth, shape, and other boundaries are not differentiated. Some choices for the disparity measure at the pixel resolution ($\sigma = 1$) are discussed in [49]. Here we shall only consider the case $\sigma > 1$. Each $\langle s,t \rangle_\sigma$ is then associated with two adjacent blocks of pixels D_1, D_2, of equal size and shape, and straddling the elementary boundary segment. Then for the raw data we take $\Delta_{s,t}(y) = \nu(y(D_1), y(D_2))$ whereas for texture analysis we utilize the transforms $y^{(1)}, \ldots, y^{(m)}$ described earlier and $\Delta_{s,t}(y)$ is given by (5.8). Thus, $\Phi_{s,t} > 0$ if and only if $\nu(y^{(i)}(D_1), y^{(i)}(D_2)) \ge c_i \beta$ for *some* transform i.

The selection of J is straightforward because our expectations about boundaries are easily expressed by local contraints. Consider the patterns

$$\begin{array}{ccc} & 0 & \\ 0 & 1 & 0 \end{array} \qquad \begin{array}{ccc} & 1 & \\ 1 & 0 & 1 \end{array} \qquad \begin{array}{cc} 1 & 1 \\ 1 & 1 \end{array} \qquad \begin{array}{ccc} & 1 & \\ 1 & 1 & 1 \\ & 1 & \end{array} \tag{5.11}$$

together with rotations through $k\pi/2$, $k = 1, 2, 3$. These can be associated with any resolution by the obvious scaling, and correspond, respectively, to an isolated or abandoned segment, a sharp turn, "small structure", and junction. Depending on σ, the pixel resolution, and available scene information, we may or may not include the latter three. In any event, $J(x)$ is simply the number of occurrences in x of the chosen patterns. In particular, there are no individual parameters for "weighting" boundary configurations, as in §4; in effect, the model is $H + \infty V$.

In principle, the estimators $\widehat{x}_{map}$, $\widehat{x}_{samp}$ can be computed by constrained simulated annealing and constrained stochastic relaxation. We have found $\widehat{x}_{samp}$ quite reliable; recall this means a sample from the conditional law

$$\Pi_T(x|y) = \frac{1}{\widehat{z}_T} 1_{\{J=0\}} \exp\left[- \sum_{\langle s,t \rangle_\sigma} (1 - x_s x_t)\, \phi(\Delta_{s,t}(y)) \right]$$

The choice of temperature T is important; obviously if T is not small enough the detected boundaries will be somewhat chaotic, whereas if T is too small the relaxation time (= # sweeps until equilibrium is reached) becomes unacceptable (see Grenander [68]). Despite the results in §3, we do not follow the logarithmic schedule for $\lambda_k \uparrow \infty$, but rather allow

λ_k to grow *linearly*. Moreover, in some of the experiments reported in [49], we did not seek a true sample, but instead employed the ICM algorithm of Besag [9], in which, in effect, $T = 0$. Indeed, we were able to segment all but the most difficult texture mosaics with this procedure. Recall that when the set of sites A_{k+1} is visited for updating, we defined $X(k+1)$ by replacing the coordinates of $X(k)$ in A_{k+1} by a *sample* drawn from the conditional distribution

$$\Pi_{T_{k+1},\lambda_{k+1}}(x_s,\, s \in A_{k+1} \mid x_s = X_s(k),\, s \notin A_{k+1},\, y)$$

The zero-temperature version replaces the sample by the *mode*, i.e. the most likely vector $(x_s,\, s \in A_{k+1})$ given the data and $(x_s = X_s(k),\, s \notin A_{k+1})$. This generates a deterministic sequence $X(k), k = 0, 1, 2 \ldots$ depending only on $X(0), \Pi$, and $\{A_k\}$. During the k^{th} sweep of the lattice S_σ^L, with $\lambda \equiv \lambda_k$, the energy $H(x,y) + \lambda_k J(x)$ is successively reduced, just as in ICM. However, since λ_k is varying, there is no fixed (or reference) energy and the interpretation as iterative improvement is lost. In Figures 5,6 we show the results of several experiments. The first one, at resolution $\sigma = 5$, uses the Kolmogorov distance (5.6) on the raw grey levels, and was obtained with the zero-temperature version of constrained stochastic relaxation; the latter two, of sizes 256×256 and 512×512, employ the measure based on (5.8) at resolutions $\sigma = 5$ and $\sigma = 8$ respectively, and were obtained by sampling at a fixed "low" temperature. All use the penalties above and block updates, with each A_k consisting of a cross of five sites. Other details and experiments may be found in [49].

5.4.3 Partitions

The label sites S_σ^L are defined exactly as in (5.9), except that here $S^L = S_1^L$ is a copy of S^P. The maximum number of distinct regions is P, which must be specified in advance. A partition is then a mapping $x: S_\sigma^L \to \{0, 1, \ldots, P-1\}$ and the regions are simply the sets $x^{-1}(k) \subset S_\sigma^L$. (It is then obvious how to label *all* the pixels if so desired.) Again, larger σ's yield more reliable results but lose boundary accuracy.

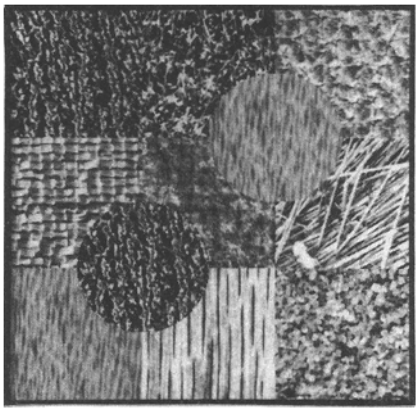
The neighborhood system in S_σ^L is *non-local* because local ones lead to deep local minima and even "spurious" global minima, as we shall see. Let $\langle s,t \rangle_\sigma$ denote a neighbor pair. Each label site $s \in S^L$ is associated with a block of pixels D_s centered at s; these blocks may be overlapping when their centers are nearby. Referring to (5.2), $\Psi_{s,t}(x) = 1_{\{x_s = x_t\}}$ and, in the case of first-order statistics,

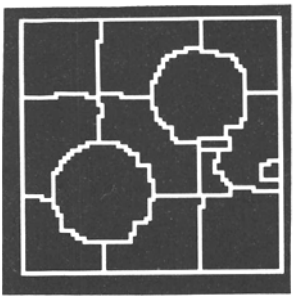
$$\Phi_{s,t}(y) = 2\, 1_{\{\nu(y(D_s),\, y(D_t)) > c\}}(y) - 1$$

Thus, $\Phi_{s,t} = +1$ or -1 depending on whether $\nu > c$ or $\nu \le c$. In this way, $\Phi_{s,t} = 1$ corresponds to dissimilar blocks and is coupled with distinct labels $(x_s \neq x_t)$ and $\Phi_{s,t} = -1$ corresponds to similar blocks and identical labels $(x_s = x_t)$. For higher-order statistics, with $\Gamma_1, \ldots, \Gamma_m$ as above,

$$\Phi_{s,t}(y) = \max_{1 \le i \le m} \left[2\, 1_{\{\nu(y^{(i)}(D_s),\, y^{(i)}(D_t)) > c_i\}}(y) - 1 \right]$$

Returning to the choice of the neighborhood system $\mathcal{G}_s$ in S_σ^L, a few simple examples will highlight the issues. Suppose the data y consists of R *constant* regions and we seek a

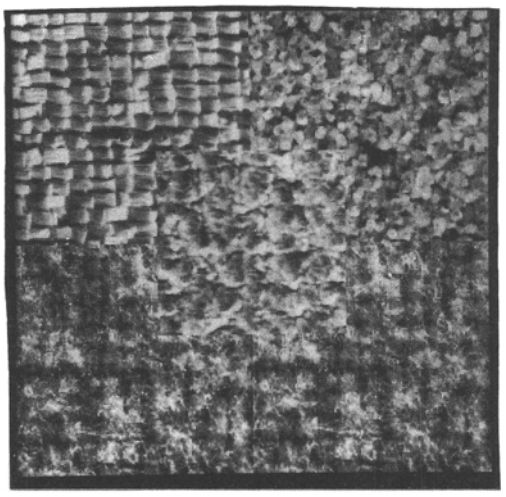


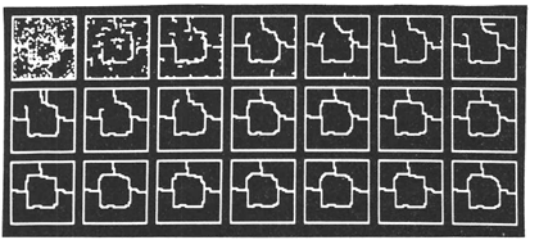


RESOLUTION $\sigma = 5$

ICM ALGORITHM

10 ITERATIONS





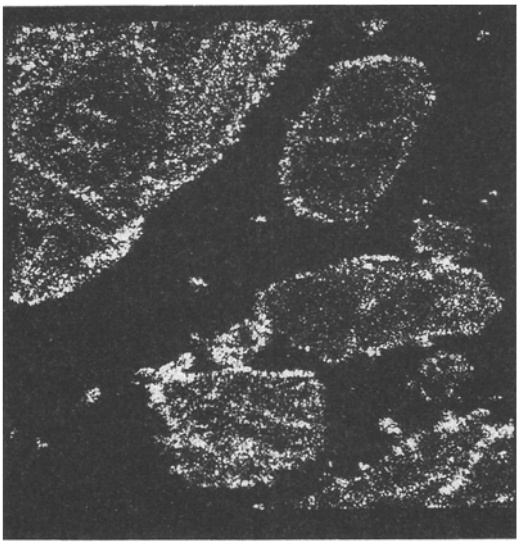
RESOLUTION $\sigma = 5$

LOW TEMPERATURE SAMPLING

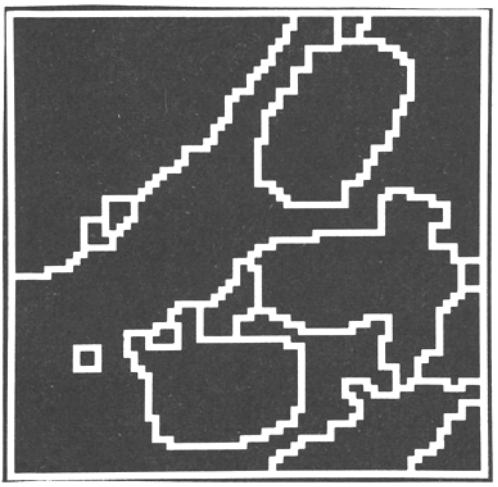
60 ITERATIONS; EVERY

3rd ITERATION SHOWN

FIGURE 5



SYNTHETIC APERTURE
RADAR IMAGE OF
SEA AND ICE (512 × 512)



RESOLUTION $\sigma = 8$
LOW TEMPERATURE SAMPLING

FIGURE 6

labeling (in this case a redundant one) on the full lattice $S^L = S^P$. The obvious disparity measure to choose is $\Phi_{s,t}(y) = -1$ if $y_s = y_t$ and $\Phi_{s,t}(y) = 1$ if $y_s \neq y_t$. Hence

$$H(x,y) = \sum_{\langle s,t\rangle_1} 1_{\{x_s=x_t\}}(1_{\{y_s\neq y_t\}} - 1_{\{y_s=y_t\}}). \tag{5.12}$$

Suppose $\mathcal{G}_s$ consisted only of sites close to s, say the four nearest ones. Suppose further that $R = 2$ (so that $y_s \in \{0,1\}$), $P = 3$, and that region "0" ($\{s: y_s = 0\}$) is split into two disjoint pieces by region "1" ($\{s: y_s = 1\}$). Then the function in (5.12) has two kinds of global minima: "correct labelings", using two of the three possible labels, and "spurious labelings", in which all three labels are used and the two parts of region "0" are interpreted as distinct. Similarly, if $R = 3, P = 3$, and y consists of three regions but two of these have no common border, then again H is minimized by both correct labelings, using all three labels, and spurious ones, using but two labels, and hence merging two of the regions. There is yet another difficulty with local neighborhoods: the existence of "wide" local minima. The most familiar example is the symmetric Ising model with no external field, in which case it is virtually impossible for a *local* Monte Carlo relaxation algorithm to evolve to one of the two global minima (all $+1$ or all -1) by starting from a configuration with two very large patches. Indeed, the energy "landscape" of H in (5.12) for $R = 1, P = 2$ is *identical* to that of the Ising model and the same is true for the case $R = 2, P = 2$, although this is less obvious: the two Ising minima may be identified with the two correct labelings $x_s = 0 \Longleftrightarrow y_s = 0$ and $x_s = 1 \Longleftrightarrow y_s = 0$ by a suitable transformation.

One remedy for these problems, namely label ambiguities and severe local minima, is to introduce long range interactions. Indeed, in the examples above, just a few of these would eliminate the "spurious" global minimum. The *distance* $\mu(s,t)$ between two sites of a graph is the smallest number of edges that must be crossed to travel between the sites s, t. In the large graph limit, *random* graphs have the minimum *diameter* ($= \sup_{s,t} \mu(s,t)$) among all graphs of a fixed degree, meaning that $|\mathcal{G}_s| \equiv$ constant. On the other hand, local graphs have large average distances: This is a problem because local relaxation algorithms evolve towards a global equilibrium by propagating relationships from site to site via the edges of the graph.

In view of these heuristics, we have chosen to include *randomly* chosen edges in $\mathcal{G}_s$. More specifically, we take $|\mathcal{G}_s| = 20$, with sixteen random neighbors for s as well as the four "near" neighbors, meaning the closest two horizontal and vertical neighbors whose associated pixel blocks do not overlap. (For example, for $\sigma = 7$ and using 21×21 blocks, the near neighbors have two intervening sites in S^L_σ.) These near neighbors are included because disparity data is *not* perfect (as in the examples above) and the local edges tend to bond nearby sites and effectively increase the number of long range interactions. The effect is dramatic; for example, in a series of experiments with "perfect" disparity data we could always achieve the global minimum by *single site iterative improvement. In fact, we conjecture that for potentials such as those appearing in (5.12), there are no local minima over random graphs in the appropriate large graph limit.*

Finally, the function J is chosen to exclude two types of label arrays: those with very small regions and those with very narrow regions. The actual implementation requires specifying the corresponding scale parameters and we refer the reader to [49] for the details.

One experiment from [49] is shown in Figure 7. The image is 216×216 and the label lattice S_σ^L is 16×16. The partitions were randomly initiated by choosing the labels independently and uniformly from $0, 1, \ldots, P-1$. Thereafter, the sites were visited and updated one at a time, using the zero-temperature sampling scheme with λ_k increasing linearly with the number of sweeps through the label array. The resolution was $\sigma = 13$ and the disparity measure was constructed from the raw intensity data together with four data transformations: an isotropic residual, two directional residuals, and one based on the range over a small window; see [49]. The most important parameters are the thresholds $c_i, 1 \leq i \leq 5$, associated with the Kolmogorov-Smirnov statistics. These were chosen by examining the histograms of the Kolmogorov-Smirnov distances for block pairs within homogeneous samples of the textures. Thresholds were set such that no more than three or four percent of these intra-region distances would be above threshold ("false alarms"). Conceivably, with enough transforms, one could set conservative (high) and nearly universal thresholds, and be assured that visibly distinct textures would be segmented with respect to *at least one* of the transforms.

This experiment illustrates a *hazard* of long range bonds. There is a gradual but marked lighting variation across the top of the original image (see Figure 7). Consequently, for the raw intensity data, the Kolmogorov-Smirnov distance between blocks from the left and right sides is large. The unreliability of the raw data results in a large value for c_1 and the partitioning is based primarily on the four transforms, for which the corresponding thresholds are much smaller.

6 Assorted Issues and Open Problems

6.1 Parameter Estimation

This is a hazardous problem, arising in virtually every application (in particular all the preceding ones) and regarded by some as a serious drawback to the methodology, especially if these models are to be put to practical use. In part, the problem stems from the complexity of the likelihood function, $\Pi(y;\theta)$, where y denotes the observed data, possibly a *sequence* of images, representing multiple observations or infinite-graph approximations. Other complicating factors include the dimensionality of the data, phase transitions, and especially the loss of information associated with the existence of unobserved variables and corrupted measurements.

Consider the simplest case: multiplicative parameters (i.e. an *exponential family*) and complete observations. Thus

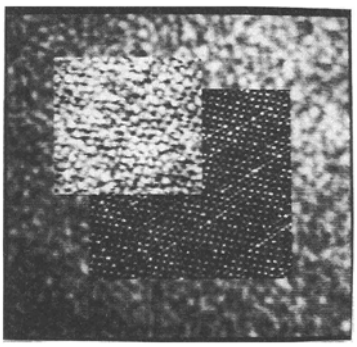
$$\Pi(x;\theta) = \exp\left\{-\sum_{j=1}^{J} \theta_j H_j(x)\right\} / Z(\theta)$$

and x is fully observed, i.e. $y = x$. Applying the basic identity

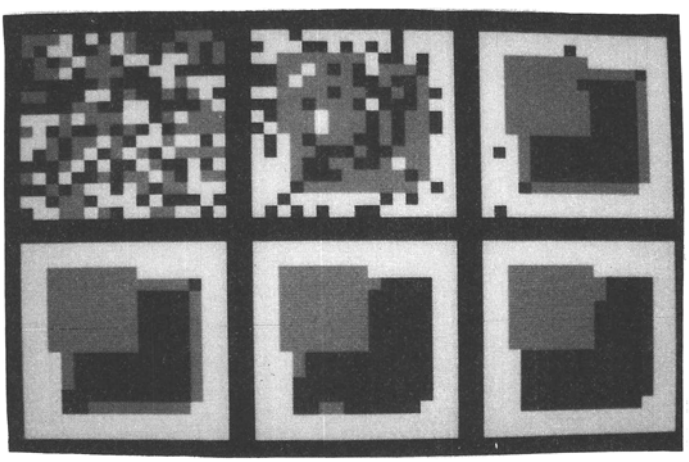
$$\frac{\partial}{\partial \theta_j} \log Z(\theta) = -E_\theta H_j(X)$$

we arrive at the familiar likelihood equations

$$E_\theta H_j = H_j(y), \quad 1 \leq j \leq J \tag{6.1}$$



ORIGINAL IMAGE
(RUG, PLASTIC, CLOTH)



SEGMENTATION ; EVERY
FIFTH SWEEP WITH
RANDOM INITIALIZATION

FIGURE 7

The groundwork for lattice-based systems and Gibbs measures may be found in Besag [7] (Gaussian case), Pickard [108,109] (for the Ising model) and in Gidas [57], including treatments of strong consistency in the infinite graph limit and asymptotic normality. Comets and Gidas [29] study the asymptotics of the ML estimators for the Curie-Weiss model: $H_1 = \sum x_s, H_2 = (\sum x_s)^2$. Younes [130,131] develops a stochastic gradient algorithm for convergence to the ML estimate in which the new estimate $\widehat{\theta}_{n+1}$ depends on both the current estimate $\widehat{\theta}_n$ and the current value X^{n+1} of a Markov chain generated with the Gibbs sampler using the evolving estimates of θ, i.e. the Markov chain with transitions $\Pi_{a_n}(x_{a_n}|y_{(a_n)})$ based on $\widehat{\theta}_n$. In a similar vein, Lippman [97] uses stochastic relaxation to estimate Lagrange multipliers associated with maximum entropy extensions from mean energy constraints; these multipliers may then be interpreted as maximum likelihood estimates. Examples of unresolved questions concerning ML estimation with fully observed data include finding optimal conditions under which the ML estimates are asymptotically normal and efficient, and *non-parametric* ML estimation, regarding the Gibbs measures as parametrized by interaction potentials from an infinite-dimensional space.

An appealing alternative to ML estimation for many cases of fully observed data (e.g. in texture synthesis) is Besag's [8] *maximum pseudo-likelihood* (MPL) in which a "likelihood" $L(\theta; x)$ is constructed from the *local* characteristics of the field:

$$L(\theta; x) = \prod_{s \in S} P_\theta(X_s = x_s \mid X_{(s)} = x_{(s)})$$

It is easy to check that for multiplicative parameters the function L is generally a *strictly concave* function of θ and that the system of equations $\nabla \log L = 0$ involves only expectations with respect to the local characteristics. An early application of a closely related approach ("coding" [11]) appears in Cross and Jain [30] in the context of texture synthesis. S. Geman and C. Graffigne [52] established consistency in the large graph limit under mild conditions; see also Guyon [70]. Chalmond [19] studied the particular case of Gaussian priors and additive white noise. Recently, Comets [28] has investigated consistency and optimality using large deviation methods. Open issues for MPL estimators include asymptotic normality (for which phase transitions appear to be relevant) and relative asymptotic efficiency between MPL and ML estimation when both are asymptotically normal.

For exponential families with incomplete data, it suffices to consider the case of a *projection*: the joint distribution of X and Y is

$$\Pi(x, y; \theta) = \exp\left\{ -\sum_{j=1}^{J} \theta_j H_j(x, y) \right\} / Z(\theta)$$

but only Y is observed. Then

$$\frac{\partial}{\partial \theta_j}(-\log \Pi(y; \theta)) = \frac{\partial}{\partial \theta_j} \log Z(\theta) + \frac{\sum_x H_j(x, y) \exp\left\{ -\sum_{j=1}^{J} \theta_j H_j(x, y) \right\}}{\sum_x \exp\left\{ -\sum_{j=1}^{J} \theta_j H_j(x, y) \right\}}$$

from which it follows that the incomplete data likelihood equations are

$$E_\theta H_j(X, Y) = E_\theta(H_j(X, y)|y), \quad 1 \le j \le J \tag{6.2}$$

Consider for instance the common case in which the posterior energy consists of "prior" components $\sum \theta_j H_j(x)$ and a data component, say $D_y(x)$, involving a *known* (multiplicative) noise parameter. Estimation of the θ_j is then based on the system

$$E_\theta H_j(X) = E_\theta(H_j(X)|y), \quad 1 \le j \le J \tag{6.3}$$

Difficulties with (6.2) or (6.3) stem from the non-concavity of the incomplete data likelihood $\log \Pi(y;\theta)$ and iterative methods such as the EM algorithm are difficult to execute. Consider for instance (6.3) with $J = 1$, and assume that $\theta \to E_\theta H(x)$ is strictly monotone and known (or pre-computed); then the M (=maximization) step is trivial, but there still remains the problem of computing or estimating the global and data-dependent expectations on the righthand side of (6.3) at each instance of the E (=expectation) step. Some of the theoretical issues are addressed in Gidas [58] and a specific example of (6.3) for a single "smoothing parameter" may be found in Geman and McClure [54]. Chalmond [20] considers the problem of restoring images corrupted by signal-dependent noise assuming both the noise and MRF parameters are unknown, and develops an iterative procedure ("Gibbsian EM algorithm") for doing the parameter estimation and the image reconstruction *at the same time*. Simultaneous estimation-reconstruction schemes were also proposed in Lakshmanan and Derin [96] and D. Geman [45].

There are by now numerous *problem-specific* methods in which simple and sometimes efficient estimators can be constructed by exploiting specific model structures, for instance particular noise mechanisms, prior models, or simplifying assumptions on the number of grey levels or graph structure. Examples include Frigessi-Piccioni [41,42], Acuna [1], Devijver and Dekesel [38], Graffigne [63], Marroquin, Mitter and Poggio [100], Ripley [115], S. Geman and McClure [54], Dinten, Guyon and Yao [31], Derin, Elliott, and Kuang [33], and Possolo [112]. The least-squares method proposed in the latter two references can be quite powerful in special circumstances. Recently Almeida and Gidas ([2,3]) introduced a variational method for estimating parameters for both complete and incomplete data.

Two fundamental open problems are to find a general but practical method for handling incomplete data and nonparametric estimation of the prior model.

6.2 Stochastic Relaxation

The notation in that of §3; in particular, $\Pi(x) \propto \exp(-H(x))$ is a Gibbs measure over $\Omega = \Lambda^S$, where $\Lambda = \{0, 1, \ldots, L\}$ and the elements of S are denoted $\{1, 2, \ldots, N\}$.

6.2.1 Simulation

We begin with the binary case ($L = 1$) and consider the class $\mathcal{U}$ of all reversible dynamics (P_{xy}) in which a site is selected at random and a "flip" is entertained with some probability depending only on the energy difference $\Delta_{xy} = H(y) - H(x)$ between the current state $x \in \Omega$ and the candidate state $y \in \Omega$. Thus $(P_{xy}) \in \mathcal{U}$ if $\Pi(x)P_{xy} = \Pi(y)P_{yx}$ and

$$P_{xy} = Q_{xy} g(H(y) - H(x)), \quad y \ne x$$

for some $g: \mathbf{R} \to [0, 1]$, where Q_{xy} is the uniform distribution over all $y \in \Omega$ with Hamming distance 1 from x: $Q_{xy} = N^{-1} 1_{\Omega_x}(y)$, with $\Omega_x = \{y \in \Omega: y_j \ne x_j \text{ for exactly one } j\}$. In

particular, both the standard Metropolis (§3.2.1) and Gibbs Sampler (§3.2.3) dynamics belong to $\mathcal{U}$. More specifically, it is easy to see that g must satisfy the condition

$$g(\Delta_{xy}) = e^{-\Delta_{xy}} g(-\Delta_{xy})$$

and it follows (cf. J. Horowitz, private communication) that the transitions $\mathcal{P} \in \mathcal{U}$ are in one-to-one correspondence with the function class $\mathcal{F} = \{g: [0,\infty) \to [0,1]: g(\Delta) \leq e^{-\Delta}\}$, since the restriction to $[0,\infty)$ of any g as above must belong to $\mathcal{F}$ and, conversely, any $g \in \mathcal{F}$ is extended to a function on $(-\infty,\infty)$ by defining $g(\Delta) = e^{-\Delta} g(-\Delta)$ for $\Delta \leq 0$.

The Metropolis dynamics corresponds to $g_M(\Delta) = e^{-\Delta^+}$ and the Gibbs Sampler to $g_S(\Delta) = (1+e^{\Delta})^{-1}$. Since g_M achieves the upper bounds on *both* half-lines (i.e. $g_M(\Delta) \equiv 1$ for $\Delta \leq 0$ and $g_M(\Delta) = e^{-\Delta}$ for $\Delta \geq 0$), we see that the *Metropolis dynamics may be characterized as the one most likely to change.* However, this property may not correspond to a rapid convergence to equilibrium: $P^n_{xy} \to \Pi(y)$. To entertain such comparisons, let us fix some family of measures $\Pi(x)$, for instance the symmetric Ising models

$$H(x;\beta) = -\beta \sum_{\langle i,j \rangle} (2x_i - 1)(2x_j - 1), \quad \beta \geq 0$$

where the lattice sites i and j may be the usual nearest-neighbors or perhaps more distant. *For which dynamics $\mathcal{P} \in \mathcal{U}$ is the rate of convergence to equilibrium most rapid?* The natural criterion here is the magnitude of the second eigenvalue of $\mathcal{P}$, although some other characterization could be of interest. For the particular example above, we would expect faster convergence for small β with g_S than with g_M (or perhaps all g) since for $\beta = 0$ (the i.i.d. case), the g_S-dynamics reaches equilibrium as soon as every site has been updated at least once. Similar questions may be formulated for more complex energy classes, e.g. those containing a data-dependent component corresponding, say, to symmetric channel noise (§2.3.1).

In another direction, we might lift the restriction to binary systems but limit the class of dynamics. Specifically, consider the four possibilities corresponding to random (R) or deterministic (D) site visitation and the Metropolis (M) or Gibbs (G) sampling mechanism. Here (D) refers to one full sweep through S. Thus,

$$(RM) \qquad P_{xy} = \frac{1}{NL} 1_{\Omega_x}(y) \exp(-[H(y) - H(x)]^+)$$

$$(RG) \qquad P_{xy} = \frac{1}{N} \sum_{k=1}^{N} 1_{x_{(k)}=y_{(k)}} \Pi_k(y_k | x_{(k)})$$

$$(DM) \qquad P = \prod_{k=1}^{N} P^k, \; P^k_{xy} = \frac{1}{L} 1_{x_{(k)}=y_{(k)}} \exp(-[H(y) - H(x)]^+)$$

$$(DG) \qquad P = \prod_{k=1}^{N} P^k, \; P^k_{xy} = 1_{x_{(k)}=y_{(k)}} \Pi_k(y_k | x_{(k)})$$

Again, for a specific class of Π's and based on the rate of convergence to equilibrium, one might compare (RM) with (RG) or (DM) with (DG). Moreover, if the Markov chains

associated with (RM) and (RG) are only observed at times $t = kN, k = 0, 1, 2, \ldots$, one might compare these convergence rates with (DM) and (DG), respectively. Some results along these lines are now available; see Hwang and Sheu [86].

Finally, how do relaxation times scale with the graph size? For example, let $\Pi_N(x) \propto \exp \sum_{\langle i,j \rangle} V(x_i, x_j)$ for some pair potential V, fix one dynamics, say the Gibbs Sampler, and let $(X(k))$ denote the corresponding Markov chain. Then for any starting measure μ, standard estimates yield

$$\sum_{x \in \Omega} |P_\mu(X(kN) = x) - \Pi_N(x)| \sim c_N(\delta_N)^k$$

for some constants $c_N > 0, 0 < \delta_N < 1$. Under what conditions is $\sup_N \delta_N < 1$, i.e., the relaxation time is actually *independent* of the number of sites? Glauber [59] has partial results (in continuous time) for special one-dimensional binary systems, and Holley [78] proves that the relaxation time is indeed independent of N for a class of one-dimensional Ising models. Nonetheless, all the major questions appear to remain open in two dimensions.

6.2.2 Annealing.

Let $\Omega_{\min} = \{x \in \Omega: H(x) = \min_y H(y)\}$ as in §3.3, and recall that under appropriate conditions on the rate of decrease of temperatures $T_n \searrow 0$, we obtain

$$\lim_{k \to \infty} P(X(k) \in \Omega_{\min} \mid X(0) = y) = 1 \text{ for all } y \in \Omega$$

where $(X(k))$ denotes the Markov chain with transitions $P_{xy}^k = Q_{xy} \exp(-[H(y) - H(x)]^+ / T_k)$ (Metropolis) or $P_{xy}^k = 1_{y_{(k)} = x_{(k)}} \Pi_{k,T_k}(y_k | x_{(k)})$ (Gibbs Sampler), where $\{\Pi_{k,T}\}$ are the local characteristics of $\Pi_T \propto \exp(-H/T)$. Once again, the relative rates of convergence is an open question and it would be informative to compare asymptotic expansions of the second eigenvalues of the (time-dependent) transition matrices.

The logarithmic annealing schedule is certainly slow and we might ask how much is lost by using faster schedules. For instance, for the geometric schedule $T_n = \gamma^n, 0 < \gamma < 1$, what is the behavior of $k \to P(X(k) \in \Omega_{\min})$ or of the deviation of $P(X(k) \in dx)$ from $\Pi_0(dx)$, the uniform measure on $\Omega_{\min}$?

From a still more practical viewpoint, what about *finite-time annealing*? Given H, etc., what is the "best" annealing schedule (T_m) over a fixed number of updates, say $1 \le m \le M$? Specifically, for example, suppose we select $X(0)$ at random and seek to minimize $E[\min_{1 \le m \le M} H(X(m))]$ over all schedules $(T_m, 1 \le m \le M)$. Are there optimal schedules which are possibly non-monotonic?

Finally, many of the same problems may be formulated for *continuous-time* annealing. For continuous Λ, the basic process is $dX(t) = -\nabla H(X(t))dt + \sqrt{2T(t)}dW_t$, whereas for discrete Λ, one constructs the Markov chain with transitions of the form $P_{xy}(t) = Q_{xy} \exp(-\alpha_{xy}/T(t))$ for $Q_{xy} > 0$. (The asymptotic behavior is the same if Q varies regularly with time, say $c_1 Q_{xy} \le Q_{xy}(t) \le c_2 Q_{xy}$, for some fixed Q and $c_1, c_2 > 0$).

6.2.3 Recent work.

Some remarkable and state-of-the-art results have been obtained by the group at Academia Sinica, Tapei, including T.-S. Chiang, Y. Chow, C.-R. Hwang, and S.-J. Sheu. For exam-

ple, there are now results on comparative dynamics and on the expected hitting time to global minima. Some of this recent work is listed in the references.

6.3 Prior Models

Ideally, the choice of the "prior" distribution $\Pi(x)$ would be largely determined by specific information about the image attributes under study. For example, in classification problems in which individual pixels are given symbolic labels, we might know the size of the actual regions or the types of boundaries between regions. Similarly, in a reconstruction problem, we might know the original surfaces are (approximately) planar or quadric. Unfortunately, however, this sort of prior knowledge is usually not available, and only generic assumptions are reliable, in which case, as noted by many authors, the choice of Π is generally guided by experience and trial-and-error rather than by formal methods of non-parametric estimation or procedures for model-fitting such as those in time-series analysis.

Here is a concrete example to illustrate the issues. Suppose we seek to recover the original distribution x of some physical quantity, such as radiant energy or isotope concentration, from imperfect measurements, y. For example, the transformation from x to y may involve blur and noise (§4) or an attenuated Radon transform (§2.3.3). In any case, the degradation mechanism is modeled as a probability distribution $\Pi(y|x)$. Let us suppose further that each component x_s assumes integer values over a relatively large dynamic range $0, \ldots, L$ (modifications for continuous values are straightforward) and that we *know* that the "true" distribution x^* is locally *constant*, or approximately so. This is an idealized world, to be sure, but all the remarks below apply as well to increasingly richer worlds by replacing the first-order differences $x_s - x_t$ by higher-order differences constructed from gradients, Laplacians, and so forth. Specifically, then, consider the family of prior models

$$\Pi(x) \propto exp - (\beta \sum_{<s,t>} \phi((x_s - x_t)/\delta))$$

where $< s,t >$ denotes a neighbor pair, say nearest-neighbor, δ is a scaling constant which depends on the dynamic range and will henceforth be ignored, β is the usual "regularization parameter", and ϕ is even and non-decreasing on $(0, \infty)$, say $\phi(0) = 0$. We have already discussed the problem of fixing ϕ and estimating β, whereas here we are concerned with the problem of *choosing* ϕ, and treating β, like δ, as a nuisance parameter.

The quadratic "potential" $\phi(u) = u^2$ is inappropriate for any imagery containing sharp transitions because large intensity gradients are severely penalized; indeed, two regions may be no more "different" at, say, two hundred grey levels apart than at fifty grey levels apart. On the other hand, there are strong computational advantages in having ϕ be at least *convex*. Estimates are based on the posterior distribution which is usually of the form

$$\Pi(x|y) \propto \exp\left[-\beta \sum \phi\left(\frac{x_s - x_t}{\delta}\right) - D_y(x)\right]$$

where $D_y(x)$ is the energy function corresponding to the degradation model, i.e. $D_y(x) = -\log \Pi(y|x) +$ const. In addition, $D_y(x)$ is often *convex* in x for each y; for example, for blur and Gaussian noise, $D_y(x) \propto \|y - \mathcal{K}x\|^2$ where $\mathcal{K}$ is the blur operator. Consequently, if ϕ is also convex, then computations based on $\Pi(x|y)$ are considerably simplified. These

observations have prompted some researchers to adopt non-quadratic, convex ϕ's, such as $\phi(u) = \log(\cosh u)$ (Green [65]), and $\phi(u) = |u|$ (Besag [12]) or $\phi(u) = u^2, |u| \leq \alpha, = 2\alpha|u| - \alpha^2, |u| > \alpha$. The potential $\phi(u) = |u|$ has the appealing property that the marginal mode is the *median* of the neighbor values. Others have noted that *finite* asymptotic behavior ($\phi(\infty) < \infty$) is more compatible with the existence of sharp transitions between regions. Thus, for example, one finds priors ([54]) with ϕ of the form

$$\phi(u) = |u|^\gamma/(1 + |u|^\gamma), \gamma > 0. \tag{6.4}$$

If ϕ is concave (e.g. $\gamma \leq 1$ in (6.4)), then the potential is strictly *non-interpolating* in the sense that, conditioned on differing intensities at nearby pixels and ignoring the data term, the most likely transition is a pure step edge, whereas if ϕ is not concave (e.g. $\gamma > 1$) there is a greater tendency to absorb a transition over a range of pixels, i.e. to interpolate across boundaries. In any case, these non-convex potentials lead to far more difficult computational problems, and there is much to be learned about this type of image modeling, particularly when higher-order smoothness conditions are imposed.

6.4 Performance Criteria.

We have seen that the same problem (de-blurring, segmentation, etc.) may attract diverse approaches, corresponding to varying choices for image models, estimators, and algorithms. This has been the tendency in the field at large and the absence of performance criteria has been repeatedly mentioned (e.g. in image analysis surveys) but rarely addressed.

Image modeling and computational issues were discussed in the previous sections, and we shall not pursue these matters any further. Consequently, we will conclude with a few remarks, within the framework of these lectures, about the choice of *estimators*. The reader is referred to Ripley [115,116], Besag [10], and Marroquin et al [100] for further commentary.

Let $\hat{x}$ denote an estimator of an image attribute x based on measurements y; we shall assume $\Pi(x|y)$ is given. Many natural, but ad hoc, estimators have been proposed, and we shall mention several of these below; for now, however, we shall restrict ourselves to those derived from loss functions $L: \Omega \times \Omega \to [0, \infty)$. Let $\hat{x}_L$ be the function $f: \mathbf{R}^M \to \Omega$ which minimizes $EL(X, f(Y))$ where M is the dimension of y; the estimate is then $\hat{x}_L(y)$. Familiar examples are zero-one loss ($L(x, \hat{x}) = 1_{\{x \neq \hat{x}\}}$), squared-error loss ($L(x, \hat{x}) = \sum_s (x_s - \hat{x}_s)^2$), and misclassification rate ($L(x, \hat{x}) = \sum_s 1_{\{x_s \neq \hat{x}_s\}}$). The corresponding estimators are $\hat{x}_{map} = \arg\max_x \Pi(x|y), \hat{x}_{mmse} = E(x|y)$, and $\hat{x}_{mpm} = \{\hat{x}_s\}, \hat{x}_s = \arg\max_{x_s} \Pi(x_s|y)$, where mpm stands for "marginal posterior mode" (or "maximizer of posterior marginals" [100]).

The appropriate loss function is necessarily problem-specific. For image restoration, reconstruction and classification, there is considerable doubt about the utility of the MAP estimator. Some authors (e.g. [10], [38], [100]) conclude this estimator often leads to gross mislabelings, over-smoothing and the obliteration of small structures. Support for this claim may be found in the work of Greig, Porteus, and Seheult [66], in which the *exact* MAP estimator is computed for certain binary systems using a network flow algorithm. Indeed, the efficacy (and visual appearance) of the MAP estimator degrades rapidly as

the relative influence in $\Pi(x|y)$ of the prior distribution (the Ising model in [66]) is increased with respect to the data component; see also Dinten, Guyon, and Yao [31]. This phenomenon appears to be related to the existence of *phase transitions* in the prior model: reconstructions may be dominated by the global properties of the prior, such as long-range order (see Example 2.3), which are characteristic of certain parameter ranges. Thus "decisions" at one image location may influence those at distant and unrelated image areas; see also the experiments in Marroquin et al [100] and Ripley [115]. Nonetheless, MAP estimation has been a fundamental tool in the engineering literature and has been employed with success in many of the cited references. A central issue is then the extent to which order phenomena in the prior are inherited by the posterior. Basically, the alternative proposed has been either $\widehat{x}_{mpm}$, for classification problems or reconstructions over a small dynamic range, and $\widehat{x}_{mmse}$, for restoration/reconstruction problems with continuous or nearly continuous levels. Notice that, unlike $\widehat{x}_{map}$, the definitions of $\widehat{x}_{mpm}$ and $\widehat{x}_{mmse}$ depend on the scale factor (= temperature) in the posterior distribution.

Pixel-based error measures are usually too local for certain other classification problems, such as boundary classification. Thus, if $x = \{x_{\langle r,t\rangle}\}$ denotes a binary boundary map indexed by the edges of the pixel lattice, then the $\widehat{x}_{mpm}$ estimator is $\widehat{x}_{\langle r,t\rangle} = 1$ if $P(x_{\langle r,t\rangle} = 1|y) > \frac{1}{2}$ and $\widehat{x}_{\langle r,t\rangle} = 0$ if $P(x_{\langle r,t\rangle} = 1|y) \leq \frac{1}{2}$. The $\widehat{x}_{mpm}$ estimator then lacks the fine structure we expect of boundary maps because placement decisions are made *individually* (based on the data) and pending decisions at nearby locations are not explicitly considered. Thus, the segmentation studies in Geman et al [49] are based on approximations to the MAP estimator.

Another possibility is to design other loss functions. For example, for image segmentation or restoration we might choose

$$L(x, \widehat{x}) = \sum_i 1_{x_{A_i} \neq \widehat{x}_{A_i}}$$

where (A_i) denotes a partition of S, say into $k \times k$ squares. The Bayes estimator, call it $\widehat{x}_{map(k)}$, is then determined by the multi-dimensional marginal modes; that is, $(\widehat{x}_s)$ must satisfy $\Pi(\widehat{x}_s, s \in A|y) = \max_{x_{A_i}} \Pi(x_{A_i}|y)$ for each A_i. We might also allow the regions to overlap; but the resulting estimator is no longer easy to compute. Notice that zero-one loss and misclassification rate are special cases corresponding to the coarsest $((A_i) = S)$ and finest $(A_i = \{s_i\})$ partitions. Ideally, if prior information were available about the scales of "important" structures, then we could choose k accordingly; for example, we might select k such that there are no significant structures of order $k \times k$ or smaller. Another possibility is to incur a loss whenever errors *aggregate*, since given a fixed number of errors, the reconstructed or labeled image may appear more faithful to the original if these errors are scattered rather than accumulated into potential artifacts. Consequently, we might choose

$$L(x, \widehat{x}) = \sum_s 1_{\{x_s \neq \widehat{x}_s\}} + \lambda \sum_{\langle s,t\rangle} 1_{\{x_s \neq \widehat{x}_s, x_t \neq \widehat{x}_t\}}$$

so that penalty λ occurs whenever two neighboring pixels are both misclassified. Analogous loss functions for reconstruction problems over a large dynamic range are then obvious. The problem is that the corresponding estimator is very complex due to the fact that each pixel appears in several terms. Finally, loss functions for multiple image attributes are discussed in [98],[100]; for example, for simultaneous edge detection and

surface interpolation, one may construct a composite function based on squared-error loss (pixels) and misclassification rate (edges).

Certain estimators not derived from loss functions have been frequently used with considerable success. The ICM algorithm (§2.3.4, 5.4.2), which corresponds to coordinate-wise minimization of the (posterior) energy, uses only the local properties of the Markov random field and is easy to implement. A variation of this algorithm, replacing the marginal mode with the marginal mean, and referred to as ICE (for iterated conditional expectations) has recently been explored for tomographic reconstruction (cf. S. Geman, private communication). Lastly, just a *sample* from the posterior distribution at "low" temperature may be effective for certain problems, such as texture segmentation (§5.4.2).

A Imaging Systems

A.1 Introduction

Objects emit radiant energy because they are illuminated or are themselves a primary source of energy. The energy emitted passes through some medium, such as the atmosphere or the human body, is intercepted by an image formation system, and a two-dimensional brightness pattern is generated on an "image plane." In a perfect system (i.e., one with no scattering, noise, or other distortions), this brightness pattern would correspond to an "ideal" intensity distribution $f(u)$.

Whereas this process is basically a continuum phenomenon, the actual recorded data, the *digital image*, consists of a finite set of measurements $g(s), s \in S$, where S is a two-dimensional rectangular lattice whose elements are referred to as *pixels* for "picture elements". Typically, S has dimension $2^n \times 2^n$ where n ranges from about six to twelve (in satellite images). In addition, the values assumed by g are *quantized* to a fixed set of integer values, for example the set $\{0, 1, \ldots, 255\}$ in "8-bit quantization."

This appendix is primarily concerned with the transformation from f to g. This involves the "degradation" of f by optical blurring, scattering, radiometric distortion and various types of noise (e.g., quantum, thermal, and film grain) as well as by the process of discretization itself, in particular digitization and quantization. The classical image restoration problem is to recover the ideal distribution f from the recorded values g, and is considered in an example in §2.3.3 for the special case of emitted energy encountered in single photon emission tomography and in §5 for general systems.

(**Note:** We shall not be concerned with *coherent* imaging (e.g. holography and synthetic aperture radar) and the analysis of *phase information*, but rather only with *incoherent* imaging with sensors which respond only to the intensity of light at selected frequencies.)

Surely the most familiar imaging systems are optical ones involving visible light and ordinary optical instruments such as lenses and cameras. We refer the reader to Horn [80] for an introduction, in the context of image analysis, of such issues as the geometric correspondence between the three-dimensional object coordinates and the two-dimensional image coordinates, different kinds of lenses, and especially the problem of the loss of *depth* information.

Penetrating radiation systems constitute another general imaging modality; for example, certain types of radiant energy, such as very short electromagnetic waves (gamma

rays, etc.) and ionized, high-energy particles, can penetrate through matter opaque to ordinary light. Another example is nuclear magnetic resonance, in which electro-magnetic energy is selectively absorbed within the body and then re-emitted and detected by the imaging device. There are basically two situations. "Active" objects emit radiation and one is usually interested in determining the internal distribution of energy from the intercepted energy; for example, in single photon emission tomography, a radioactive material is placed inside the body and escaping photons are detected and counted by sensors arranged outside the body. Usually, these sensors involve an array of parallel bore collimators through which the photons pass; depending on the thickness of the plate and size of the apertures, there is a trade-off between background interference from scatter and low counts.

In the other case, there is a source of radiation directed at the object of interest; some emitted particles pass through the object and are intercepted. The objects are "passive", whereas the imaging system itself is "active"; an example is ordinary hospital radiology. An image is formed because the recorded intensity differs from the emitted intensity due to attenuation. The most familiar cases are biomedical, but another example occurs in industrial x-ray systems in which scattering becomes significant. In general, attenuation and scattering are complex phenomena, depending on the radiation wavelength, the atomic composition of the objects, etc. Many of the standard texts on image processing provide more details, for instance [4],[60],[113].

Finally, "brightness" and "intensity" are usually informal terms used in place of more formal concepts. For example, the energy flux emitted from a point on a three-dimensional surface at a particular instant is called the *radiance* and is measured in units of power per area per steradian; see e.g. [80]. It consists of two components, *illumination*, a property of the energy source, and *reflectance*, a property of the surface materials. The energy flux incident on the image plane is referred to as *irradiance* and measured in power per unit area (e.g., watts per square meter). Both radiance and irradiance are then functions of time, position, wavelength, and other factors, and are in fact proportional to each other under ideal conditions and after the appropriate transformation has been made between object and image coordinates.

A.2 Point Spread Functions

Blurring at the image formation stage is common. For example, in an optical imaging system with ordinary lenses, the pattern formed on the image plane will not generally correspond to the actual radiance pattern unless the system is properly focused. In addition, there is often some distortion or loss of spatial resolution due to other factors, such as motion and scattering. As a result, the brightness pattern actually formed, say $b(u), u \in \mathbf{R}^2$, will generally differ from the true pattern $f(u)$, depending on the medium, sensor, etc.

In simple imaging systems, the transformation $b = \Phi(f)$ is generally assumed to be *linear* and *space- (or shift-) invariant*, which means that if f is translated by v, say $f_v(u) = f(u - v)$, then b is correspondingly shifted and $\Phi(f_v) = b_v$. These conditions amount to saying that Φ is given by a convolution in $\mathbf{R}^2$:

$$\Phi(f)(v) = \iint f(u)K(v-u)du$$

for some function K, referred to as the *point spread function* (PSF) of the imaging system, and interpreted as the response of the system to an infinite point source of light (or "impulse") at the origin. Usually, $K \geq 0$ and has total mass one in order to conserve the total "energy" of the distribution f. If the PSF is unknown, it can sometimes be estimated by examining the blur patterns produced by placing point sources at various locations in the image plane.

The Fourier transform $\hat{K}$ of K,

$$\hat{K}(\xi) = \iint e^{-i\xi \cdot u} K(u) du$$

is called the *modulation transfer function*, and blurring acts as a multiplicative filter in the frequency domain: $\hat{b} = \hat{K}\hat{f}$. Gaussian PSF's result from diverse phenomena, for example long-term exposure to atmospheric turbulence, in which case

$$K(u) = (2\pi\sigma)^{-1} exp - (|u|^2/2\sigma^2) \;\; \hat{K}(\xi) = exp(-|\xi|^2\sigma^2/2)$$

where σ measures the degree of blurring or, equivalently, of attenuation of the high-frequency components. Another example of a "low-pass filter" is a de-focused (circular) lens:

$$K(u) = \begin{cases} (\pi R^2)^{-1} & \text{if } |u| \leq R \\ 0 & \text{otherwise} \end{cases}$$

where R depends on the diameter of the lens and other factors. Finally, a moving object (or camera) will generate "motion blur." For constant velocity, this results in a modulation transfer function of the form $\hat{K}(\xi) = sin(l\omega)/l\omega)$ where $\omega = \xi \cdot (cos\theta, sin\theta)$, θ being the direction of motion and $2l$ the blur extent.

The formation process may be neither linear nor space-invariant, nor even independent of the particular energy distribution f. Perhaps the most general formulation encountered is

$$\Phi(f)(v) = \iint K(v, u, f(u)) du$$

where K may also depend on time and wavelength. (The linear case is $K(v, u, z) = K(v, u)z$, and the space-invariant case is $K(v, u, z) = K(u - v, 0, z)$.) For instance, in certain industrial inspection devices, dense objects are subjected to high-energy x-rays, resulting in PSF's of the form $K(v, u, z) = (2\pi z)^{-1} exp - \{|u - v|^2/2z^2\}$.

A.3 Sensor Distortion and Noise

The brightness distribution b formed on the image plane is detected by a sensing device, introducing a transformation of b sometimes referred to as *radiometric distortion* (or *sensor nonlinearities*). Moreover, the sensing and recording processes also introduce various types of noise and other artifacts. For simplicity we shall assume the entire process occurs in a continuum; in reality, the image plane is finite and some discretization may be inherent in the detection phase, converting b into a discrete-parameter energy distribution. In addition, we shall concentrate on the special case of photoelectronic systems, although most sensors are based on the same principles and the overall situation is essentially the same for photochemical and other technologies. The case of tomographic imaging is considered in §2.3.3.

Sensors are equipped with a photoactive or photosensitive surface which responds to incoming photons by releasing charge carriers, which are collected and measured, and which determine an "electron image." However, the situation is complicated by several sources of noise. First, there is inherent randomness in the number of photons striking the surface; these quantum fluctuations are accurately modeled as a Poisson random measure, approximately Gaussian for large counts. Secondly, there is not a one-to-one correspondence between the number of incident photons and the number of charge carriers which are released. Some photons are deflected or lost, and some charge may be randomly added. Roughly, the ratio of photon to charge flux is called the *quantum efficiency* and depends on the wavelength, photoactive material, etc. Some devices have high efficiency (e.g. solid state) and others relatively low efficiency (e.g. film and vacuum devices).

The accuracy of measurement of the incident photon activity, N, over an area ∂A over a time interval ∂t is gauged by the *signal-to-noise ratio* $E(N)/(Var(N))^{1/2}$. Since the predominant noise effect is Poisson for such sensors, this reduces to $(E(N))^{1/2}$. It is therefore difficult to obtain high signal-to-noise ratios (say 1000) without increasing ∂A and/or ∂t, which results in a loss of spatial and/or temporal resolution. In other words, there is a trade-off between localization and measurement accuracy. Finally, there is *thermal noise* in the electronics, which is characteristically assumed to be spatially uncorrelated and Gaussian.

The electron image is measured by a scanning device. The current r in the scanning beam is proportional to the number of electrons emitted during the readout period. In the absence of noise, the relationship between the measured current and incident light intensity is approximated by $r = Cb^\gamma$ for some positive γ called the *gamma* of the detector. Incorporating the effects of noise, the actual amount of current in the readout beam is

$$g(u) = C(b(u))^\gamma + C^{1/2}(b(u))^{\gamma/2}\eta_Q(u) + \eta_t(u)$$

where η_t denotes (Gaussian) thermal noise and η_Q denotes (approximately Gaussian) quantum noise. A common simplification is to replace $(b(u))^{\gamma/2}$ by an *average* over the image plane; this is the assumption of "signal-independent noise."

A.4 Sampling and Quantization

If the intensities recorded by the sensing device are still in continuous form (for example an electrical signal representing catenated scan lines), then it must be sampled over a *finite* set S. This "digitization" process may be at least partly inherent in the sensor hardware (e.g. charge coupled devices have a square array of sensors which limits the possible resolution) or, as in the case of film, may involve a separate device, such as a rotating drum scanner. The *digital image* is then $g(s), s \in S$. We regard $g(s)$ as the intensity of radiant energy *detected* by the imaging system in a small spatial-temporal cell $\partial A_s \times \partial t$; the actual energy flux is different, as already discussed. Usually, S is a rectangular array; for instance T.V. frames (single images) have dimension 450x560. For simplicity, we shall always assume S is a square lattice: $S = \{(i,j)|1 \leq i,j \leq N\}$. It should be noted that some loss of information is unavoidable due to the *diffraction limit* which results from the wave nature of light and is the ultimate resolution limit of any system. In particular, it makes no sense to have pixels corresponding to areas smaller than the order of the wavelength of light.

The values assumed by g are quantized to a finite set of values referred to as *grey levels* or *greytones*. Usually, g assumes integer values $0 \leq k \leq 2^m - 1$; for example, in "8-bit quantization", $g \in \{0, \ldots, 255\}$. There is some evidence that humans do not perceive an appreciable deterioration with fewer than eight bits, perhaps even six; see Gonzalez and Wintz [60].

It is interesting to ask how much information is lost by the process of digitization itself. Intuitively, at least, it would seem that sampling would "lose" the high frequency components corresponding to visual detail. For convenience, assume $f \in L^1(\mathbf{R})$ with Fourier transform

$$\hat{f}(u) = \int e^{-2\pi i u x} f(x) dx$$

and suppose that f is *band-limited*, meaning that $\hat{f}(u) = 0$ for $|u| > w$ for some constant w. Let $\hat{f}_E$ denote the periodic extension of $\hat{f}$ with period $2w$ and expand this extension in a Fourier series:

$$\hat{f}_E(u) = \sum_{k=-\infty}^{\infty} a_k e^{-2\pi i (uk/2w)}$$

Hence,

$$\begin{aligned} f(x) &= \int \hat{f}(u) e^{2\pi i u x} du \\ &= \int_{-w}^{w} \hat{f}_E e^{2\pi i u x} du \\ &= \sum_k a_k \int_{-w}^{w} e^{-iu((\pi k/w) - 2\pi x)} du \\ &= 2w \sum_k a_k sinc(\pi(k - 2xw)) \end{aligned}$$

where $\mathrm{sinc}(u) = u^{-1} \sin u$. Substituting $x = k/2w$ into the equation above, it follows that

$$f(x) = \sum_k f(k/2w) sinc(\pi(k - 2xw))$$

Therefore, f is recovered from its "sampled values" $f(k/2w), -\infty < k < \infty$. The extension to two-dimensions is the

Whittaker-Shannon Sampling Theorem Let f and $\hat{f}$ be a Fourier transform pair in $\mathbf{R}^2$, and suppose $\hat{f}$ is band-limited, say $\hat{f} = 0$ for $|u_1| > w_1, |u_2| > w_2$. Then f is determined by the sampled values $\{f(k\delta_x, j\delta_y), -\infty < k, j < +\infty\}$ provided that $\delta_x \leq (1/2w_1), \delta_y \leq (1/2w_2)$.

There are some practical limitations. In particular, since the number of samples actually available is always *finite*, there will be frequency distortions in the reconstruction of f (a formula similar to the one-dimensional case) unless f also happens to be periodic. Not surprisingly, artifacts are introduced if the same reconstruction formula is applied to a non band-limited function. Finally, there are problems in matching the sensor spacings and the conditions of the theorem; see e.g. Horn [80].

REFERENCES

1. C. Acuna, (1988). "Parameter estimation for stochastic texture models", Ph.D. thesis, Department of Mathematics and Statistics, University of Massachusetts.

2. M. Almeida and B. Gidas, (1989). "A variational method for estimating the parameters of a MRF from complete or incomplete data", preprint, Brown University.

3. M. Almeida, (1989). "Statistical inference for MRF with unbounded spins and applications to texture representation", Ph.D. Thesis, Division of Applied Mathematics, Brown University.

4. H. C. Andrews and B. R. Hunt, (1977). *Digital Image Restoration.* Englewood Cliffs, NJ: Prentice-Hall.

5. R. Azencott, (1988). "Simulated Annealing", *Seminaire Bourbaki*, 1987-88, no. 697.

6. P. Baldi, (1984). "Stochastic destabilisation of local minima", Technical Report, Univ. of Pisa.

7. J. Besag, (1975). "On the estimation and testing of spatial interaction in Gaussian lattice processes", *Biometrika*, 62, pp. 555-562.

8. J. Besag and P. A. P. Moran, (1977). "Efficiency of pseudo-likelihood estimation for simple Gaussian fields", *Biometrika*, 64, pp. 616–618.

9. J. Besag, (1983). Discussion of invited papers. *Bull. Internat. Statist. Inst.*, pp. 422-425.

10. J. Besag, (1986). "On the statistical analysis of dirty pictures",(with discussion), *J. Royal Statist. Soc.*, Ser. B, 48, pp. 259–302.

11. J. Besag, (1974). "Spatial interaction and the statistical analysis of lattice systems", *J. Royal Statist. Soc.*, Ser. B, 36, pp. 192–236.

12. J. Besag, (1989). "Towards Bayesian image analysis," *J. Appl. Statistics*, 16.

13. A. Blake, (1983). "The least disturbance principle and weak constraints", *Pattern Recognition Letters*, 1, pp. 393–399.

14. T. C. Brown, C. Jennison, and B. W. Silverman, (1987). "Edge process models for regular and irregular pixels", preprint.

15. T. M. Cannon, H. J. Trussell, and B. R. Hunt, (1978). "Comparison of image restoration methods", *Applied Optics*, 17, pp. 3385–3390.

16. J. Canny, (1986). "A computational approach to edge detection", *IEEE Trans. Pattern Anal. Machine Intell.*, 8, pp. 679-698.

17. O. Catoni, (1988). "Optimal cooling schedules for annealing", *C.R. Acad. Sci. Paris*

18. V. Cerny, (1985). "Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm", *J. Opt. Theory Appl.*, 45, pp. 41-51.

19. B. Chalmond, (1988). "Image restoration using an estimated Markov model", *Signal Processing*, 15, pp. 115-129.

20. B. Chalmond, (1988). "An iterative Gibbsian technique for reconstruction of m-ary images", *Pattern Recognition*, to appear.

21. R. Chellappa, T. Simchony and H. Jinchi, (1988). "Relaxation algorithms for MAP restoration of gray level images with multiplicative noise", Technical Report, Signal and Image Processing Institute, Univ. Southern Calif.

22. T.-S. Chiang and Y. Chow, (1987). "On eigenvalues and optimal annealing rate", Technical Report, Institute of Math., Academia Sinica, Taipei, Taiwan.

23. T.-S. Chiang and Y. Chow, (1987). "On the convergence rate of annealing processes", to appear in *SIAM J. Control and Optimization.*

24. T.-S. Chiang and Y. Chow, (1987). "A limit theorem for a class of inhomogeneous Markov processes", Technical Report, Institute of Math., Academia Sinica, Taipei, Taiwan.

25. T.-S. Chiang and Y. Chow, (1989). "The asymptotic behavior of simulated annealing processes with absorption", Technical Report, Institute of Math., Academia Sinica.

26. T.-S. Chiang, C.-R Hwang, S.-J Sheu, (1987). "Diffusion for global optimization in R^n", *SIAM J. Control and Optimization*, 25, pp. 737–753.

27. F. S. Cohen and D. B. Cooper, (1987). "Simple parallel hierarchical and relaxation algorithms for segmenting noncausal Markovian random fields", *IEEE Trans. Pattern Anal. Machine Intell.*, 9, pp. 195–219.

28. F. Comets, (1989). "On the consistency of a class of estimators for exponential families of Markov random fields on the lattice", preprint, Universite de Paris-X.

29. F. Comets and B. Gidas, (1988). "Asymptotics of maximum likelihood estimators for the Curie-Weiss model", Technical Report, Div. of App. Math., Brown University.

30. G. R. Cross and A. K. Jain, (1983). "Markov random field texture models", *IEEE Trans. Pattern Analysis and Machine Intelligence*, 5, pp. 25-39.

31. J. M. Dinten, X. Guyon, and J. F. Yao, (1988). "On the choice of the regularization parameter: the case of binary images in the Bayesian restoration framework", preprint, Universite de Paris-Sud.

32. J. M. Dinten, (1988). "Tomographic reconstruction with a limited number of projections: regularization using a Markov model", preprint, Universite de Paris-Sud.

33. H. Derin, H. Elliott, and J. Kuang, (1985). "A new approach to parameter estimation for Gibbs random fields", preprint, University of Massachusetts.

34. H. Derin and W. S. Cole, (1986). "Segmentation of textured images using Gibbs random fields", *Comput. Vision, Graphics, and Image Process.*, 35, pp. 72–98.

35. H. Derin and H. Elliott, (1987). "Modelling and segmentation of noisy and textured images using Gibbs random fields", *IEEE Trans. Pattern Anal. Machine Intell.*, 9, pp.

39–55.

36. H. Derin, H. Elliott, R. Cristi, and D. Geman, (1984). "Bayes smoothing algorithms for segmentation of binary images modeled by Markov random fields", *IEEE Trans. Pattern Analysis and Machine Intelligence*, 6, pp. 707–720.

37. H. Derin and C. Won, (1986). "A parallel image segmentation algorithm using relaxation with varying neighborhoods and its mapping to array processors", Tech. Report, Dept. of Elec. and Comp. Engineering, University of Massachusetts.

38. P. A. Devijver and M. M. Dekesel, (1987). "Learning the parameters of a hidden Markov random field image model: a simple example", in *Pattern Recognition Theory and Applications*, P. A. Devijver and J. Kittler, Eds., Heidelberg: Springer-Verlag, pp. 141–163.

39. J. R. Erhman, L. D. Fosdick, and D. C. Handscomb, (1960). "Computation of order parameters in an Ising lattice by the Monte Carlo method," *J. Mathematical Phys.*, 1, pp. 547-558.

40. B. R. Frieden, (1972). "Restoring with maximum likelihood and maximum entropy", *J. Opt. Soc. Amer.*, 62, pp. 511–518.

41. A. Frigessi and M. Piccioni, (1988). "Consistent parameter estimation for 2-D Ising fields corrupted by noise: numerical experiments", Quaderno, IAC-CNR, Roma.

42. A. Frigessi and M. Piccioni, (1988). "Parameter estimation for the two-dimensional Ising fields corrupted by noise", Quaderno, IAC-CNR, Roma.

43. A. Gagalowicz and Song De Ma, (1985). "Sequential synthesis of natural textures", *Comput. Vision, Graphics, Image Process.*, 30, pp. 289–315.

44. E. Gamble and T. Poggio, (1987). "Visual integration and detection of discontinuities: the key role of intensity edges", A.I.Memo No. 970, M.I.T. A.I. Laboratory.

45. D. Geman, (1985). "Bayesian image analysis by adaptive annealing", in *Digest 1985 Int. Geosci. Remote Sensing Symp.*, IGARSS '85, Amherst.

46. D. Geman, (1987). "A stochastic model for boundary detection", *Image and Vision Computing*, 5, pp. 61-65.

47. D. Geman and S. Geman, (1987). "Relaxation and annealing with constraints", *Complex Systems Technical Report No. 35*, Div.of Applied Mathematics, Brown University.

48. D. Geman, S. Geman, and C. Graffigne, (1987). "Locating texture and object boundaries", in *Pattern Recognition Theory and Applications*, P. A. Devijver and J. Kittler, Eds., Heidelberg: Springer-Verlag.

49. D. Geman, S. Geman, C. Graffigne, and P. Dong, (1987). "Boundary detection by constrained optimization", Technical Report, Dept. of Mathematics, Univ. of Massachusetts.

50. S. Geman and D. Geman, (1984). "Stochastic relaxation, Gibbs distributions, and

the Bayesian restoration of images", *IEEE Trans. Pattern Anal. Machine Intell.*, 6, pp. 721–741.

51. S. Geman, (1983). "Stochastic relaxation methods for image restoration and expert systems", presented at *ARO Workshop: Unsupervised Image Analysis*, Brown Univ.; appeared in *Maximum Entropy and Bayesian Methods in Science and Engineering*, Vol.1, G.J.Erickson and C.R.Smith, eds. Kluwer Academic Publishers, 1988.

52. S. Geman and C. Graffigne, (1986). "Markov random field image models and their applications to computer vision", in: *Proceedings of the International Congress of Mathematicians, 1986.* Ed. A. M. Gleason, American Mathematical Society, Providence.

53. S. Geman and C.-R. Hwang, (1986). "Diffusion for global optimization", *SIAM J. Control and Optimiation* vol. 24, pp. 1031–1043.

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

55. B. Gidas, (1985). "Nonstationary Markov chains and convergence of the annealing algorithm", *J. Stat. Phys.*, 39, pp. 73–131.

56. B. Gidas, (1989). "A renormalization group approach to image processing problems", *IEEE Trans. Pattern Anal. Machine Intell.*, 11, pp. 164-180.

57. B. Gidas, (1988). "Parameter estimation for Gibbs distributions, I: Fully observed case", preprint, Brown University.

58. B. Gidas, "Parameter estimation for Gibbs distributions, II: Partially observed data", in preparation.

59. R. Glauber, (1963). "Time-dependent statistics of the Ising model", *J. Math. Physics*, 4, pp. 294-307.

60. R. C. Gonzalez and P. Wintz, (1977). *Digital Image Processing.* Reading:Addison-Wesley.

61. J. Goutsias and J. M. Mendel, (1987). "Semi-Markov random field models for image segmentation", IEEE Inter. Conf. Acoustics, Speech, Signal Proc., Dallas, 1987.

62. J. Goutsias and J. M. Mendel, (1987). "Semi-Markov random field models for texture synthesis", SPIE Inter.Sympos. Pattern Recog.and Acoustical Imaging.

63. C. Graffigne, (1987). "Experiments in texture analysis and segmentation", Ph.D. Dissertation, Div. of Applied Mathematics, Brown University.

64. P. J. Green, (1986). Discussion: "On the statistical analysis of dirty pictures" by Julian Besag, *J. of the Royal Stat. Society*, B-48, pp. 259–302.

65. P. J. Green, (1989). "Bayesian reconstruction from emission tomography data using a modified EM algorithm," *IEEE Trans. Medical Imaging*, to appear.

66. D. M. Greig, B. T. Porteous, and A. H. Seheult, (1989). "Exact M.A.P. estimation

for binary images", *J. Royal Statist. Soc.*, Ser. B.

67. U. Grenander, (1976,78,81). *Lectures in Pattern Theory*, vols. I–III. New York: Springer-Verlag.

68. U. Grenander, (1983). "Tutorial in Pattern Theory", Technical Report, Div. of Applied Mathematics, Brown University.

69. D. Griffeath, (1976). "Introduction to random fields", in *Denumerable Markov Chains*, Kemeny, Knapp and Snell, Eds. New York: Springer-Verlag.

70. X. Guyon, (1986). "Estimation d'un champ de Gibbs", Universite de Paris - I, preprint.

71. A. Habibi, (1972). "Two-dimensional Bayesian estimate of images", *Proc. IEEE*, 60, pp. 878–883.

72. B. Hajek, (1985). "Cooling schedules for optimal annealing", preprint.

73. B. Hajek, (1985). "A tutorial survey of theory and applications of simulated annealing", *Proceedings 24the IEEE Conf. Decision and Control* pp. 755–760.

74. J. M. Hammersley and D. C. Handscomb, (1964). *Monte Carlo Methods.* London: Methuen and Company.

75. J. M. Hammersley and P. Clifford, (1968). "Markov fields of finite graphs and lattices", Univ. of Calif.-Berkeley, preprint.

76. M. Hassner and J. Sklansky, (1980). "The use of Markov random fields as models of texture", *Computer Graphics and Image Processing*, 12, pp. 357–370.

77. W. K. Hastings, (1970). "Monte Carlo sampling methods using Markov chains and their applications", *Biometrika*, 57, pp. 97-109.

78. R. Holley, (1985). "Rapid convergence to equilibrium in one-dimensional stochastic Ising models", *Ann. Probability*, 13, pp. 72-89.

79. R. Holley and D. Stroock, (1987). "Simulated annealing via Sobolev inequalities", preprint.

80. B. K. P. H. Horn, (1986). *Robot Vision.* Cambridge: M.I.T. Press.

81. R. Hummel, Kimia, and Zucker, (1987). "Deblurring Gaussian blur", *Comput. Vision Graphics Image Processing*, 38, pp. 66-80.

82. B. R. Hunt, (1973). "The application of constrained least-squares estimation to image restoration by digital computer", *IEEE Trans. Computers, 22*, pp. 805-812.

83. B. R. Hunt, (1977). "Bayesian methods in nonlinear digital image restoration", *IEEE Trans. Comput.*, vol. C-23, pp. 219–229.

84. C.-R. Hwang and S.-J. Sheu, (1987). "Large time behaviors of perturbed diffusion Markov processes with applications", I, II, and III. Technical Report, Institute of Math., Academia Sinica.

85. C.-R. Hwang and S.-J. Sheu, (1988). "On the weak reversibility condition in simulated annealing", Technical Report, Institute of Math., Academia Sinica.

86. C.-R. Hwang and S.-J. Sheu, (1989). "Remarks on Gibbs sampler and Metropolis sampler", Technical Report, Institute of Math., Academia Sinica.

87. C.-R. Hwang and S.-J. Sheu, (1988). "Singular perturbed Markov chains and the exact behaviors of simulated annealing processes, Technical Report, Institute of Math., Academia Sinica.

88. E. Ising, (1925). *Zeitschrift Physik*, 31, p. 253.

89. D. Isaacson and R. Madsen, (1976). *Markov Chains: Theory and Applications.* New York: John Wiley and Sons.

90. F.-C. Jeng and J. Woods, (1988). "Compound Gauss-Markov random fields for image estimation", Technical Report, Rensselaer Polytechnic Institute.

91. F.-C. Jeng and J. Woods, (1988). "Simulated annealing in compound Gaussian random fields", Technical Report, Rensselaer Polytechnic Institute.

92. J. W. Kay, (1988). "On the choice of regularisation parameter in image restoration", *Springer Lecture Notes in Computer Science*, 301, pp. 587-596.

93. R. Kinderman and J. L. Snell, (1980). *Markov Random Fields and Their Applications.* Providence, RI: Amer. Math. Soc.

94. S. Kirkpatrick, C. D. Gellatt, Jr. and M. P. Vecchi, (1983). "Optimization by simulated annealing", *Science*, 220, pp. 671–680.

95. P. J. M. van Laarhoven and E. H. L. Aarts, (1987). *Simulated Annealing: Theory and Applications.* Dordrecht: D.Reidel.

96. S. Lakshmanan and H. Derin, (1989). "Simultaneous parameter estimation and segmentation of Gibbs random fields using simulated annealing", *IEEE Trans. Pattern Anal. Machine Intell.*, 11, pp. 799-813.

97. A. Lippman, (1986). "A maximum entropy method for expert system construction", Ph.D. Thesis, Div. Appl. Math., Brown University.

98. J. L. Marroquin, (1985). "Optimal Bayesian estimators for image segmentation and surface reconstruction", *Artifical Intell. Lab. Memo* 839, M.I.T.

99. J. L. Marroquin, (1984). "Surface reconstruction preserving discontinuities", *Artificial Intell. Lab. Memo* 792, M.I.T.

100. J. L. Marroquin, S. Mitter, and T. Poggio, (1987). "Probabilistic solution of ill-posed problems in computational vision", *J. Am. Stat. Assoc.*, 82, pp. 76–89.

101. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, (1953). "Equations of state calculations by fast computing machines", *J. Chemical Physics*, 21, pp. 1087–1091.

102. D. Mitra, F. Romeo and A. Sangiovanni-Vincentelli, (1986). "Convergence and finite time behavior of simulated annealing", *App. Prob.* vol. 18, pp. 747–771.

103. J. Moussouris, (1974). "Gibbs and Markov random systems with constraints", *J. Stat. Physics*, 10, pp. 11–33.

104. D. Mumford and J. Shah, (1986). "Boundary detection by minimizing functionals, I", preprint. See also: Proc. IEEE Conf. Comput. Vision Patt. Recog., San Francisco, 1985.

105. D. W. Murray and B. F. Buxton, (1987). "Scene segmentation from visual motion using global optimization", *IEEE Trans. Pattern Anal. Machine Intell.*, 9, pp. 220–228.

106. D. W. Murray, A. Kashko and H. Buxton, (1986). "A parallel approach to the picture restoration algorithm of Geman and Geman on an SIMD machine", *Image and Vision Computing*, 4, pp. 133-142.

107. N. E. Nahi and T. Assefi, (1972). "Bayesian recursive image estimation", *IEEE Trans. Comput.*, 21, pp. 734–738.

108. D. K. Pickard, (1979). "Asymptotic inference for Ising lattice III. Non-zero field and ferromagnetic states", *J. Appl. Prob.* 16, pp. 12–24.

109. D. K. Pickard, (1987). "Inference for discrete Markov fields: The simplest nontrivial case", *J. Amer. Statist. Assoc.*, 82, pp. 90–96.

110. M. Pincus, (1970). "A Monte-Carlo method for the approximate solution of certain types of constrained optimization problems", *Oper. Res.*, 18, pp. 1225-1228.

111. T. Poggio, V. Torre and C. Koch, (1985). "Computational vision and regularization theory", *Nature*, 317, pp. 314–319.

112. A. Possolo, (1986). "Estimation of binary Markov random fields", Technical Report, Dept. of Statistics, Univ. of Washington.

113. W. Pratt, (1978). *Digital Image Processing.* New York: John Wiley and Sons.

114. B. D. Ripley, (1977). "Modeling spatial patterns", *J. R. Statist. Soc.*, Ser. B, 39, pp. 172-212.

115. B. D. Ripley, (1988). *Statistical inference for spatial processes.* Cambridge, Cambridge University Press.

116. B. D. Ripley, (1986). "Statistics, images and pattern recognition", *Canadian J. of Statistics*, 14, pp. 83–111.

117. W. H. Richardson, (1972). "Bayesian-based iterative method of image restoration", *J. Opt. Soc. Amer.*, 62, pp. 55-59.

118. A. Rosenfeld and A. C. Kak, (1982). *Digital Picture Processing*, vols. 1, 2, 2nd ed. New York: Academic.

119. L. A. Shepp and Y. Vardi, (1982). "Maximum likelihood reconstruction in positron emission tomography", *IEEE Trans. on Medical Imaging*, 1, pp. 113-122

120. T. Simchony and R. Chellappa, (1988). "Stochastic and deterministic algorithms for MAP texture segmentation", *Proc. IEEE International Conference on Acoustics,*

Speech, and Signal Processing, 2, New York, pp. 1120–1123.

121. P. H. Swain, S. B. Vardeman, and J. C. Tilton, (1981). "Contextual classification of multispectral data", *Pattern Recognition*, 13, pp. 429–441.

122. D. Terzopoulos, (1986). "Regularization of inverse visual problems involving discontinuities", *IEEE Trans. Pattern Anal. Machine Intell.*, 8, pp. 413–424.

123. A. M. Thompson, J. C. Brown, J. W. Kay and D. W. Titterington, (1988). "A comparison of methods of choosing the smoothing parameter in image restoration by regularization", Technical Report, Dept. of Physics, Astron., and Statist., University of Glasgow.

124. D. M. Titterington, (1985). "General structure of regularization procedures in image reconstruction", *Astron. Astrophys.*, 144, pp. 381-387.

125. A. Trouve, (1988). "Synchronous simulated annealing", *C. R. Acad. Sci. Paris.*

126. H. J. Trussell, (1980). "The relationship between image restoration by the maximum a posteriori method and a maximum entropy method", *IEEE Trans. Acoust., Speech, Signal Processing*, 28, pp. 114–117.

127. J. N. Tsitsiklis, (1985). "Markov chains with rare transitions and simulated annealing", Technical Report, Laboratory for Information and Decision Sciences, Massachusetts Institute of Technology.

128. J. W. Woods, S. Dravida, and R. Mediavilla, (1987). "Image estimation using doubly stochastic Gaussian random field models", *IEEE Trans. Pattern Analysis and Machine Intelligence*, 9, pp. 245–253.

129. J.-F. Yao, (1988). "Methodes Bayesiennes en segmentation d'images", Technical Report, Universite de Paris-Sud.

130. L. Younes, (1988). "Estimation and annealing for Gibbs fields", to appear in *Annales de l'Institut Henri Poincare*, 3.

131. L. Younes, (1988). "Estimation for Gibbsian fields: applications and numerical results", Technical Report, Universite de Paris-Sud.