Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images

STUART GEMAN AND DONALD GEMAN

Abstract—We make an analogy between images and statistical mechanics systems. Pixel gray levels and the presence and orientation of edges are viewed as states of atoms or molecules in a lattice-like physical system. The assignment of an energy function in the physical system determines its Gibbs distribution. Because of the Gibbs distribution, Markov random field (MRF) equivalence, this assignment also determines an MRF image model. The energy function is a more convenient and natural mechanism for embodying picture attributes than are the local characteristics of the MRF. For a range of degradation mechanisms, including blurring, nonlinear deformations, and multiplicative or additive noise, the posterior distribution is an MRF with a structure akin to the image model. By the analogy, the posterior distribution defines another (imaginary) physical system. Gradual temperature reduction in the physical system isolates low energy states ("annealing"), or what is the same thing, the most probable states under the Gibbs distribution. The analogous operation under the posterior distribution yields the maximum *a posteriori* (MAP) estimate of the image given the degraded observations. The result is a highly parallel "relaxation" algorithm for MAP estimation. We establish convergence properties of the algorithm and we experiment with some simple pictures, for which good restorations are obtained at low signal-to-noise ratios.

Index Terms—Annealing, Gibbs distribution, image restoration, line process, MAP estimate, Markov random field, relaxation, scene modeling, spatial degradation.

I. Introduction

THE restoration of degraded images is a branch of digital picture processing, closely related to image segmentation and boundary finding, and extensively studied for its evident practical importance as well as theoretical interest. An analysis of the major applications and procedures (model-based and otherwise) through approximately 1980 may be found in [47]. There are numerous existing models (see [34]) and algorithms and the field is currently very active. Here we adopt a Bayesian approach, and introduce a "hierarchical," stochastic model for the original image, based on the *Gibbs distribution*, and a new restoration algorithm, based on stochastic relaxation and *annealing*, for computing the maximum *a posteriori* (MAP) estimate of the original image given the degraded image. This algorithm is highly parallel and exploits the equivalence between Gibbs distributions and *Markov random fields* (MRF).

Manuscript received October 7, 1983; revised June 11, 1984. This work was supported in part by ARO Contract DAAG-29-80-K-0006 and in part by the National Science Foundation under Grants MCS-83-06507 and MCS-80-02940.

S. Geman is with the Division of Applied Mathematics, Brown University, Providence, RI 02912.

D. Geman is with the Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003.

The essence of our approach to restoration is a stochastic relaxation algorithm which generates a sequence of images that converges in an appropriate sense to the MAP estimate. This sequence evolves by *local* (and potentially *parallel*) changes in pixel gray levels and in locations and orientations of boundary elements. Deterministic, iterative-improvement methods generate a sequence of images that monotonically increase the posterior distribution (our "objective function"). In contrast, stochastic relaxation permits changes that *decrease* the posterior distribution as well. These are made on a *random* basis, the effect of which is to avoid convergence to *local maxima*. This should not be confused with "probabilistic relaxation" ("relaxation labeling"), which is deterministic; see Section X.

The stochastic relaxation algorithm can be informally described as follows.

1) A local change is made in the image based upon the current values of pixels and boundary elements in the immediate "neighborhood." This change is *random*, and is generated by sampling from a local conditional probability distribution.

2) The local conditional distributions are dependent on a global control parameter T called "temperature." At *low* temperatures the local conditional distributions concentrate on states that *increase* the objective function, whereas at high temperatures the distribution is essentially uniform. The limiting cases, $T = 0$ and $T = \infty$, correspond respectively to greedy algorithms (such as gradient ascent) and undirected (i.e., "purely random") changes. (High temperatures induce a loose coupling between neighboring pixels and a chaotic appearance to the image. At low temperatures the coupling is tighter and the images appear more regular.)

3) Our image restorations avoid local maxima by beginning at high temperatures where many of the stochastic changes will actually decrease the objective function. As the relaxation proceeds, temperature is gradually lowered and the process behaves increasingly like iterative improvement. (This gradual reduction of temperature simulates "annealing," a procedure by which certain chemical systems can be driven to their low energy, highly regular, states.)

Our "annealing theorem" prescribes a schedule for lowering temperature which guarantees convergence to the global maxima of the posterior distribution. In practice, this schedule may be too slow for application, and we use it only as a guide in choosing the functional form of the temperature-time dependence. Readers familiar with Monte Carlo methods in statistical physics will recognize our stochastic relaxation algorithm as a "heat bath" version of the *Metropolis algorithm* [42]. The idea of introducing temperature and simulating an-

0162-8828/84/1100-0721$01.00

nealing is due to Černý [8] and Kirkpatrick *et al.* [40], both of whom used it for combinatorial optimization, including the traveling salesman problem. Kirkpatrick also applied it to computer design.

Since our approach is Bayesian it is model-based, with the "model" captured by the prior distribution. Our models are "hierarchical," by which we mean layered processes reflecting the type and degree of *a priori* knowledge about the class of images under study. In this paper, we regard the original image as a pair $X = (F, L)$ where F is the matrix of observable pixel intensities and L denotes a (dual) matrix of unobservable edge elements. Thus the usual gray levels are considered a marginal process. We refer to F as the *intensity process* and L as the *line process.* In future work we shall expand this model by adjoining other, mainly geometric, attribute processes.

The degradation model allows for noise, blurring, and some nonlinearities, and hence is characteristic of most photochemical and photoelectric systems. More specifically, the degraded image G is of the form $\phi(H(F)) \odot N$, where H is the blurring matrix, ϕ is a possibly nonlinear (memoryless) transformation, N is an independent noise field, and $\odot$ denotes any suitably invertible operation, such as addition or multiplication. Surprisingly, these nonlinearities do not affect the computational burden.

To pin things down, let us briefly discuss the Markovian nature of the intensity process; similar remarks apply to the line process, the pair (F, L), and the distribution of (F, L) conditional on the "data" G. Of course, all of this will be discussed in detail in the main body of the paper.

Let $Z_m = \{(i, j): 1 \leqslant i, j \leqslant m\}$ denote the $m \times m$ integer lattice; then $F = \{F_{i,j}\}$, $(i, j) \in Z_m$, denotes the gray levels of the original, digitized image. Lowercase letters will denote the values assumed by these (random) variables; thus, for example, $\{F = f\}$ stands for $\{F_{i,j} = f_{i,j}, (i, j) \in Z_m\}$. We regard F as a sample realization of a random field, usually isotropic and homogeneous, and with significant correlations well beyond nearest neighbors. Specifically, we model F as an MRF, or, what is the same (see Section IV), we assume that the probability law of F is a Gibbs *distribution.* Given a *neighborhood system* $\mathcal{F} = \{\mathcal{F}_{i,j}, (i, j) \in Z_m\}$, where $\mathcal{F}_{i,j} \subseteq Z_m$ denotes the neighbors of (i, j), an MRF over $(Z_m, \mathcal{F})$ is a stochastic process indexed by Z_m for which, for every (i, j) and every f,

$$P(F_{i,j} = f_{i,j} \mid F_{k,l} = f_{k,l}, (k, l) \neq (i, j)) = P(F_{i,j} = f_{i,j} \mid F_{k,l} = f_{k,l}, (k, l) \in \mathcal{F}_{i,j}). \quad (1.1)$$

The MRF–Gibbs equivalence provides an explicit formula for the *joint* probability distribution $P(F = f)$ in terms of an *energy function*, the choice of which, together with $\mathcal{F}$, supplies a powerful mechanism for modeling spatial continuity and other scene features.

The relaxation algorithm is designed to maximize the conditional probability distribution of (F, L) given the data $G = g$, i.e., find the mode of the *posterior distribution* $P(X = x \mid G = g)$. This form of Bayesian estimation is known as *maximum a posteriori* or MAP estimation, or sometimes as *penalized maximum likelihood* because one seeks to maximize $\log P(G = g \mid X = x) + \log P(X = x)$ as a function of x; the second term is the "penalty term." MAP estimation has been successfully employed in special settings (see, e.g., Hunt [31] and Hansen and Elliott [25]) and we share the opinion of many that the MAP formulation (and a Bayesian approach in general; see also [24], [43], [45]) is well-suited to restoration, particularly for handling general forms of spatial degradation. Moreover, the distribution of G itself need not be known, which is fortunate due to its usual complexity. On the other hand, MAP estimation clearly presents a formidable computational problem. The number of possible intensity images is L^{m^2}, where L denotes the number of allowable gray levels, which rules out any direct search, even for small ($m = 64$), binary ($L = 2$) scenes. Consequently, one is usually obliged to make simplifying assumptions about the image and degradation models as well as compromises at the computational stage. Here, the computational problem is overcome by exploiting the pivotal observation that the posterior distribution is again Gibbsian with approximately the same neighborhood system as the original image, together with a sampling method which we call the *Gibbs Sampler.* Indeed, our principal theoretical contribution is a general, practical, and mathematically coherent approach for investigating MRF's by sampling (Theorem A), and by computing modes (Theorem B) and expectations (Theorem C).

The Gibbs Sampler generates realizations from a given MRF by a "relaxation" technique akin to site-replacement algorithms in statistical physics, such as "spin-flip" and "exchange" systems. The prototype is due to Metropolis *et al.* [42]; see also [7], [18], and Section X. Cross and Jain [12] use one of these algorithms invented for studying binary alloys. ("Relaxation labeling" in the sense of [13], [30], [46], [47] is different; see Section X.) The Markov property (1.1) permits parallel updating of the line and pixel sites, each of which is "refreshed" according to a simple recipe determined by the governing distribution. Thus, both parts of the MRF–Gibbs equivalence are exploited, for computing and modeling, respectively. Moreover, minimum mean-square error (MMSE) estimation is also feasible by using the (temporal) ergodicity of the relaxation chain to compute *means* w.r.t. the posterior distribution. However, we shall not pursue this approach.

We have used a comparatively slow, raster scan-serial version of the Gibbs Sampler to generate images and restorations (see Section XIII). But the algorithm is parallel; it could be executed in essentially one-half the time with two processors running simultaneously, or in one-third the time with three, and so on. The full parallel potential is realized by assigning one (simple) processor to each site of the intensity process and to each site of the line process. Whatever the number of processors, parallel implementation is made feasible by a small communications requirement among processors. The communications burden is related to the neighborhood size of the graph associated with the image model, and herein lies much of the power of the hierarchical structure: although the field model $X = (F, L)$ has a local graph structure, the *marginal* distribution on the observable intensity process F has a *completely connected graph.* The introduction of a hierarchy dramatically expands the richness of the model of the observed process while only moderately adding to the computa-

tional burden. We shall return to these points in Sections IV and XI.

The MAP algorithm depends on an *annealing schedule*, which refers to the (sufficiently) slow decrease of a ("control") parameter T that corresponds to *temperature* in a physical system. As T decreases, samples from the posterior distribution are forced towards the minimal energy configurations; these correspond to the mode(s) of the distribution. Theorem B makes this precise, and is, to our knowledge, the first theoretical result of this nature. Roughly speaking, it says that if the temperature $T(k)$ employed in executing the kth site replacement (i.e., the kth image in the iteration scheme) satisfies the bound

$$T(k) \geqslant \frac{c}{\log (1 + k)}$$

for every k, where c is a constant independent of k, then with probability converging to one (as $k \to \infty$), the configurations generated by the algorithm will be those of minimal energy. Put another way, the algorithm generates a Markov chain which converges *in distribution* to the uniform measure over the minimal energy configurations. (It should be emphasized that *pointwise* convergence, i.e., convergence *with probability one*, is in general not possible.) These issues are discussed in Section XII, and the algorithm is demonstrated in Section XIII on a variety of degraded images. We also discuss the nature of the constant c in regard to practical convergence rates. Basically, we believe that the logarithmic rate is best possible. However, the best (i.e., smallest) value of c that we have obtained to date (see the Appendix) is far too large for computational value and our restorations are actually performed with small values of c. As yet, we do not know how to bring the theory in line with experimental results in this regard.

The role of the Gibbs (or Boltzmann) distribution, and other notions from statistical physics, in the construction of "expert systems" is expanding. To begin with, we refer the reader to [21] for the original formulation of our computational method and of a general approach to expert systems based on maximum entropy extensions. As previously mentioned, Černý [8] and Kirkpatrick *et al.* [40] introduced annealing into combinatorial optimization. Other examples include the work of Cheeseman [9] on maximum entropy and diagnosis and of Hinton and Sejnowski [29] on neural modeling of inference and learning.

This paper is organized as follows. The degradation model is described in the next section, and the undegraded image models are presented in Section IV after preliminary material on graphs and neighborhood systems in Section III. In particular, Section IV contains the definitions of MRF's, Gibbs distributions, and the equivalence theorem. Due to the plethora of Markovian models in the literature, we pause in Section V to compare ours to others, and in Section VI to explain some connections with maximum entropy methods. In Section VII we raise the issues of parameter estimation and model selection, and indicate why we are avoiding the former for the time being. The posterior distribution is computed in Section VIII and the corresponding optimization problem is addressed in Section IX. The concept of stochastic relaxation is reviewed in Section X, including its origins in physics. Sections XI and XII are devoted to the Gibbs Sampler, dealing, respectively, with its mechanical and mathematical workings. Our experimental results appear in Section XIII, followed by concluding remarks.

II. Degraded Image Model

We follow the standard modeling of the (intensity) image formation and recording processes, and refer the reader to [31] or [47] for better accounts of the physical mechanisms.

Let H denote the "blurring matrix" corresponding to a shift-invariant point-spread function. The formation of F gives rise to a blurred image $H(\mathrm{F})$ which is recorded by a sensor. The latter often involves a nonlinear transformation of $H(\mathrm{F})$, denoted here by ϕ, in addition to random sensor noise $\mathrm{N} = \{\eta_{i,j}\}$, which we assume to consist of independent, and for definiteness, Gaussian variables with mean μ and standard deviation σ.

Our methods apply to essentially arbitrary noise processes $\mathrm{N} = \{\eta_{i,j}\}$, discrete or continuous. However, computational feasibility requires that the description of N as an MRF (this can always be done; see Section IV) has an associated graph structure that is approximately "local"; the same requirement is applied to the image process $\mathrm{X} = (\mathrm{F}, \mathrm{L})$. For clarity, we forgo full generality and focus on the traditional Gaussian white noise case. Extension to a general noise process is mostly a matter of notation.

The degraded image is then a function of $\phi(H(\mathrm{F}))$ and N, say $\psi(\phi(H(\mathrm{F})), \mathrm{N})$, for example, addition or multiplication. (To compute the posterior distribution, we only need to assume that $b \to \psi(a, b)$ is invertible for each a.) For notational ease, we will write

$$\mathrm{G} = \phi(H(\mathrm{F})) \odot \mathrm{N}. \tag{2.1}$$

At the pixel level, for each $(i,j) \in Z_m$,

$$G_{i,j} = \phi\left(\sum_{(k,l)} H(i-k, j-l) F_{k,l}\right) \odot \eta_{i,j}. \tag{2.2}$$

The mathematical results require an additional assumption, namely, that F and N be independent as stochastic processes (and likewise for L and N) and we assume this henceforth. This is customary, although we recognize the limitation in certain contexts, e.g., for nuclear scan pictures.

For computational purposes, the degree of locality of F should be approximately preserved by (2.1), so that the neighborhood systems for the prior and posterior distributions on (F, L) are comparable. This is achieved when H is a simple convolution over a small window. For instance, take

$$H(k,l) = \begin{cases} \frac{1}{2}, & k = 0, l = 0 \\ \frac{1}{16}, & |k|, |l| \leqslant 1, (k,l) \neq (0,0) \end{cases} \tag{2.3}$$

so that the intensity at (i,j) is weighted equally with the average of the eight nearest neighbors. The function ϕ is unrestricted, bearing in mind that the true noise level depends on ϕ, $\odot$, and σ. Typically, ϕ is logarithmic (film) or algebraic (TV).

An important special case, which occurs in two-dimensional (2-D) signal theory, is the segmentation of noisy images into

coherent regions. The usual model is

$$G = F + N \tag{2.4}$$

where N is white noise and the number of intensity levels is small. This is the model entertained by Hansen and Elliott [25] for simple, binary MRF's F, and by many other workers with varying assumptions about F; see [14], [16], [17]. In this case, namely (2.4), we can extract simple images under extremely low signal-to-noise ratios.

The full degraded image is (G, L); that is, the "line process" is not transformed.

III. Graphs and Neighborhoods

Here and in Section IV we present the general theory of MRF's on graphs, focusing on the aspects and examples which figure in the experimental restorations. The level of abstraction is warranted by the variety of MRF's, graphs, and probability distributions simultaneously under discussion.

Let $S = \{s_1, s_2, \cdots, s_N\}$ be a set of *sites* and let $\mathcal{G} = \{\mathcal{G}_s, s \in S\}$ be a *neighborhood system* for S, meaning any collection of subsets of S for which 1) $s \notin \mathcal{G}_s$ and 2) $s \in \mathcal{G}_r \Leftrightarrow r \in \mathcal{G}_s$. Obviously, $\mathcal{G}_s$ is the set of *neighbors* of s and the pair $\{S, \mathcal{G}\}$ is a graph in the usual way. A subset $C \subseteq S$ is a *clique* if every pair of distinct sites in C are neighbors; $\mathcal{C}$ denotes the set of cliques.

The special cases below are especially relevant.

Case 1: $S = Z_m$. This is the set of pixel sites for the intensity process F; $\{s_1, s_2, \cdots, s_N\}$, $N = m^2$, is any ordering of the lattice points. We are interested in homogeneous neighborhood systems of the form

$$\mathcal{G} = \mathcal{F}_c = \{\mathcal{F}_{i,j}, (i,j) \in Z_m\};\ \mathcal{F}_{i,j}$$
$$= \{(k,l) \in Z_m : 0 < (k-i)^2 + (l-j)^2 \leqslant c\}.$$

Notice that sites at or near the boundary have fewer neighbors than interior ones; this is the so-called "free boundary" and is more natural for picture processing than torodial lattices and other periodic boundaries. Fig. 1(a), (b), (c) shows the (interior) neighborhood configurations for $c = 1, 2, 8$; $c = 1$ is the first-order or nearest-neighbor system common in physics, in which $\mathcal{F}_{i,j} = \{(i, j-1), (i, j+1), (i-1, j), (i+1, j)\}$, with adjustments at the boundaries. In each case, (i, j) is at the center, and the symbol ∘ stands for a neighboring pixel. The cliques for $c = 1$ are all subsets of Z_m of the form $\{(i, j)\}$, $\{(i, j), (i, j+1)\}$ or $\{(i, j), (i+1, j)\}$, shown in Fig. 1(d). For $c = 2$, we have the cliques in Fig. 1(d) as well as those in Fig. 1(e). Obviously, the number of clique types grows rapidly with c. However, only small cliques appear in the model for F actually employed in this paper; indeed, the degree of progress with only *pair* interactions is somewhat surprising. Nonetheless, more complex images will likely necessitate more complex energies. Our experiments (see Section XIII) suggest that much of this additional complexity can be accommodated while maintaining modest neighborhood sizes by further developing the hierarchy.

Case 2: $S = D_m$, the "dual" $m \times m$ lattice. Think of these sites as placed midway between each vertical or horizontal pair of pixels, and as representing the possible locations of "edge elements." Shown in Fig. 1(f) are six pixel sites together with seven line sites denoted by an X. The six surrounding X's are the neighbors of the middle X for the neighborhood system we denote by $\mathcal{L} = \{\mathcal{L}_d, d \in D_m\}$. Fig. 1(g) is a segment of a realization of a binary line process for which, at each line site, there may or may not be an edge element. We also consider line processes with more than two levels, corresponding to edge elements with varying orientations.

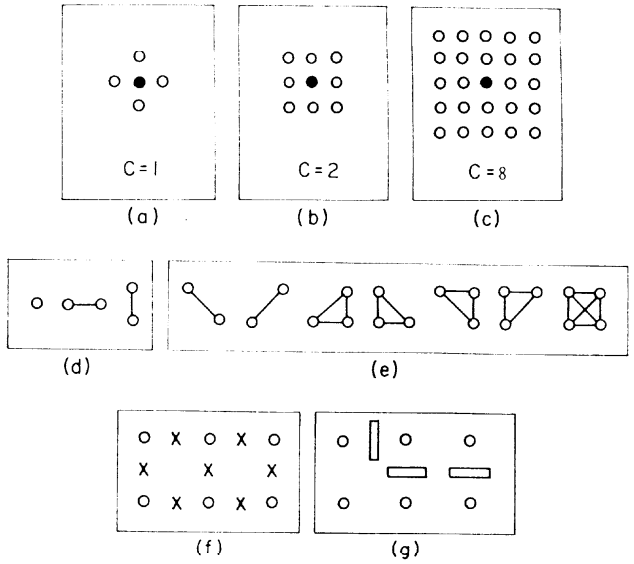


Fig. 1.

Case 3: $S = Z_m \cup D_m$. This is the setup for the field (F, L). Z_m has neighborhood system $\mathcal{F}_1$ (nearest-neighbor lattice) and D_m has the above described system. The pixel neighbors of sites in D_m are the two pixels on each side, and hence each (interior) pixel has four line site neighbors.

IV. Markov Random Fields and Gibbs Distributions

We now describe a class of stochastic processes that includes both the prior and posterior distribution on the original image. In general, this class of processes (namely, MRF's) is neither homogeneous nor isotropic, assuming the index set S has enough geometric structure to even *define* a suitable family of translations and rotations. However, the *particular* models we choose for prior distributions on the original image are in fact both homogeneous and isotropic in an appropriate sense. (This is not the case for the *posterior* distribution.) We refer the reader to Section XIII for a precise description of the prior models employed in our experiments, and in particular for specific examples of the role of the line elements.

As in Section III, $\{S, G\}$ denotes an arbitrary graph. Let $X = \{X_s, s \in S\}$ denote *any* family of random variables indexed by S. For simplicity, we can assume a common state space, say $\Lambda \doteq \{0, 1, 2, \cdots, L-1\}$, so that $X_s \in \Lambda$ for all s; the extension to site-dependent state spaces, appropriate when S consists of both line and pixel sites, is entirely straightforward (although not merely a notational matter due to the "positivity condition" below). Let Ω be the set of all possible *configurations:*

$$\Omega = \{\omega = (x_{s_1}, \cdots, x_{s_N}) : x_{s_i} \in \Lambda, 1 \leqslant i \leqslant N\}.$$

As usual, the event $\{X_{s_1} = x_{s_1}, \cdots, X_{s_N} = x_{s_N}\}$ is abbreviated $\{X = \omega\}$.

X is an MRF with respect to $\mathcal{G}$ if

$$P(X = \omega) > 0 \quad \text{for all} \quad \omega \in \Omega; \tag{4.1}$$

$$P(X_s = x_s | X_r = x_r, r \neq s) = P(X_s = x_s | X_r = x_r, r \in \mathcal{G}_s) \tag{4.2}$$

for every $s \in S$ and $(x_{s_1}, \cdots, x_{s_N}) \in \Omega$. Technically, what is meant here is that the pair $\{X, P\}$ satisfies (4.1) and (4.2) relative to some probability measure on Ω. The collection of functions on the left-hand side of (4.2) is called the *local characteristics* of the MRF and it turns out that the (joint) probability distribution $P(X = \omega)$ of *any* process satisfying (4.1) is *uniquely* determined by these conditional probabilities; see, e.g., [6, p. 195].

The concept of an MRF is essentially due to Dobrushin [15] and is one way of extending Markovian dependence from 1-D to a general setting; there are, of course, many others, some of which will be reviewed in Section V.

Notice that *any* X satisfying (4.1) is an MRF if the neighborhoods are large enough to encompass the dependencies. The utility of the concept, at least in regard to image modeling, is that priors are available with neighborhoods that are small enough to ensure feasible computational loads and yet still rich enough to model and restore interesting classes of images (and textures: [12]).

Ordinary 1-D Markov chains are MRF's relative to the nearest-neighbor system on $S = \{1, 2, \cdots, N\}$ (i.e., $\mathcal{G}_1 = \{2\}$, $\mathcal{G}_i = \{i-1, i+1\}$ $2 \leq i \leq N-1$, $\mathcal{G}_N = \{N-1\}$) if we assume all positive transitions and the chain is started in equilibrium. In other words, the "one-sided" Markov property

$$P(X_k = x_k | X_j = x_j, j \leq k-1) = P(X_k = x_k | X_{k-1} = x_{k-1})$$

and the "two-sided" Markov property

$$P(X_k = x_k | X_j = x_j, j \neq k) = P(X_k = x_k | X_j = x_j, j \in \mathcal{G}_k)$$

are equivalent. Similarly for an rth order Markov process on the line with respect to the r nearest neighbors on one side and on both sides. (This appears to be doubted in [1] but follows, eventually, from straightforward calculations or immediately from the Gibbs connection.)

Gibbs models were introduced into image modeling by Hassner and Sklansky [28], although the treatment there is mostly expository and limited to the binary case.

A *Gibbs distribution* relative to $\{S, \mathcal{G}\}$ is a probability measure π on Ω with the following representation:

$$\pi(\omega) = \frac{1}{Z} e^{-U(\omega)/T} \tag{4.3}$$

where Z and T are constants and U, called the *energy function*, is of the form

$$U(\omega) = \sum_{C \in \mathcal{C}} V_C(\omega). \tag{4.4}$$

Recall that $\mathcal{C}$ denotes the set of cliques for $\mathcal{G}$. Each V_C is a function on Ω with the property that $V_C(\omega)$ depends only on those coordinates x_s of ω for which $s \in C$. Such a family $\{V_C, C \in \mathcal{C}\}$ is called a *potential*. Z is the normalizing constant:

$$Z \doteq \sum_{\omega} e^{-U(\omega)/T} \tag{4.5}$$

and is called the *partition function*. Finally, T stands for "temperature"; for our purposes, T controls the degree of "peaking" in the "density" π. Choosing T "small" exaggerates the mode(s), making them easier to find by sampling; this is the principle of annealing, and will be applied to the posterior distribution $\pi(\boldsymbol{f}, \boldsymbol{l}) = P(\mathrm{F} = \boldsymbol{f}, \mathrm{L} = \boldsymbol{l} | \mathrm{G} = \boldsymbol{g})$ in order to find the MAP estimate. Of course, we will show that $\pi(\boldsymbol{f}, \boldsymbol{l})$ is Gibbsian and identify the energy and neighborhood system in terms of those for the priors. The *choice* of the prior distributions, i.e., of the particular functions V_C for the image model $\pi(\omega) = P(X = \omega)$, will be discussed later on; see Section VII for some general remarks and Section XIII for the particular models employed in our experiments.

The terminology obviously comes from statistical physics, wherein such measures are "equilibrium states" for physical systems, such as ferromagnets, ideal gases, and binary alloys. The V_C functions represent contributions to the total energy from external fields (singleton cliques), pair interactions (doubletons), and so forth. Most of the interest there, and in the mathematical literature, centers on the case in which S is an *infinite*, 2-D or 3-D lattice; singularities in Z may then occur at certain ("critical") temperatures and are associated with "phase transitions."

Typically, several free parameters are involved in the specification of U, and Z is then a function of those parameters—notoriously intractable. For more information see [3], [5], [6], [23], [32], and [39].

The best-known of these lattice systems is the Ising model, invented in 1925 by E. Ising [33] to help explain ferromagnetism. Here, $S = Z_m$ and $\mathcal{G} = \mathcal{F}_1$, the nearest-neighbor system.

The most general form of U is then

$$U(\omega) = \sum V_{\{i,j\}}(x_{i,j}) + \sum V_{\{(i,j),(i+1,j)\}}(x_{i,j}, x_{i+1,j}) + \sum V_{\{(i,j),(i,j+1)\}}(x_{i,j}, x_{i,j+1}) \tag{4.6}$$

where the sums extend over all $(i, j) \in Z_m$ for which the indicated cliques make sense. The Ising model is the special case of (4.6) in which X is binary ($L = 2$), homogeneous (= strictly stationary), and isotropic (= rotationally invarient):

$$U(\omega) = \alpha \sum x_{i,j} + \beta \left(\sum x_{i,j} x_{i+1,j} + \sum x_{i,j} x_{i,j+1} \right) \tag{4.7}$$

for some parameters α and β, which measure, respectively, the external field and bonding strengths.

Returning to the general formulation, recall that the local characteristics

$$\pi(x_s | x_r, r \neq s) = \frac{\pi(\omega)}{\sum_{x_s \in \Lambda} \pi(\omega)} \qquad s \in S, \omega \in \Omega$$

uniquely determine π for any probability measure π on Ω, $\pi(\omega) > 0$ for all ω. The difficulty with the MRF formulation *by itself* is that

i) the *joint* distribution of the X_s is not apparent;

ii) it is extremely difficult to spot local characteristics, i.e., to determine when a given set of functions $\psi(x_s | x_r, r \neq s)$, $s \in S$, $(x_{s_1}, \cdots, x_{s_N}) \in \Omega$, are conditional probabilities for some (necessarily unique) distribution on Ω.

For example, Chellappa and Kashyap [10] allude to i) as a disadvantage of the "conditional Markov" models. See also the discussion in [6]. In fact, these apparent limitations to the MRF formulation have been noted by a number of authors, many of whom were obviously not aware of the following theorem.

Theorem: Let $\mathcal{G}$ be a neighborhood system. Then X is an MRF with respect to $\mathcal{G}$ if and only if $\pi(\omega) = P(X = \omega)$ is a Gibbs distribution with respect to $\mathcal{G}$.

Among other benefits, this equivalence provides us with a simple, practical way of specifying MRF's, namely by specifying potentials, which is easy, instead of local characteristics, which is nearly impossible. In fact, with some experience, one can choose U's in accordance with the desired *local* behavior, at least at the intensity level. In short, the modeling and consistency problems of i) and ii) are eliminated.

Proofs may be found in many places now; see, e.g., [39] and the references therein, or the approach via the Hammersley-Clifford expansion in [6]. An influential discussion of this correspondence appears in Spitzer's work, e.g., [48]. Explicit formulas exist for obtaining U from the local characteristics. Conversely, the local characteristics of π are obtained in a straightforward way from the potentials: use the defining ratios and make the allowable cancellations. Fix $s \in S$, $\omega = (x_{s_1}, \cdots, x_{s_N}) \in \Omega$, and let ω^x denote the configuration which is x at site s and agrees with ω everywhere else. Then if $\pi(\omega) = P(X = \omega)$ is Gibbsian,

$$P(X_s = x_s | X_r = x_r, r \neq s) = Z_s^{-1} \exp - \frac{1}{T} \sum_{C: s \in C} V_C(\omega) \tag{4.8}$$

$$Z_s \doteq \sum_{x \in \Lambda} \exp - \frac{1}{T} \sum_{C: s \in C} V_C(\omega^x). \tag{4.9}$$

Notice that the right-hand side of (4.8) only depends on x_s and on x_r, $r \in \mathcal{G}_s$, since any site in a clique containing s must be a neighbor of s. These formulas will be used repeatedly to program the Gibbs Sampler for local site replacements.

For the Ising model, the conditional probability that $X_{i,j} = x_{i,j}$, given the states at $S \backslash \{i, j\}$, or equivalently, just the four nearest neighbors, reduces to

$$\frac{e^{-x_{ij}(\alpha + \beta v_{i,j})}}{1 + e^{-(\alpha + \beta v_{i,j})}}$$

where $v_{i,j} = x_{i,j-1} + x_{i-1,j} + x_{i,j+1} + x_{i+1,j}$. This is also known as the autologistic model and has been used for texture modeling in [12]. More generally, if the local characteristics are given by an exponential family and if $V_C(\omega) \equiv 0$ for $|C| > 2$, then the pair potentials always "factor" into a product of two like terms; see [6].

We conclude with some further discussion of a remark made in Section I: that the hierarchical structure introduced with the line process L expands the graph structure of the *marginal* distribution of the intensity process F. Consider first an arbitrary MRF X with respect to a graph $\{S, \mathcal{G}\}$. Fix $r \in S$ and let $\hat{X} = \{X_s, s \in S, s \neq r\}$. The marginal distribution $\hat{P}$ of $\hat{X}$ is derived from the distribution P of X by summing over the range of X_r. Use the Gibbs representation for P and perform this summation: the resulting expression for $\hat{P}$ can be put in the Gibbs form, and from this the neighborhood system on $\hat{S} \doteq S \backslash \{r\}$ can be inferred. The conclusion of this exercise is that $s_1, s_2 \in \hat{S}$ are, in general, neighbors if either i) they were neighbors in S under $\mathcal{G}$ or ii) each is a neighbor of $r \in S$ under $\mathcal{G}$. Now let X = (F, L), with neighborhood system defined at the end of Section III. Successive summations of the distribution of X over the ranges of the elements of L yields the marginal distribution of the observable intensity process F. Each summation leaves a graph structure associated with the marginal distribution of the remaining variables, and this can be related to the original neighborhood system by following the preceding discussion of the general case. It is easily seen that when all of the summations are performed, the remaining graph is completely connected; under the marginal distribution of F, all sites are neighbors. This calculation suggests that significant long-range interactions can be introduced through the development of hierarchical structures without sacrificing the computational advantages of local neighborhood systems.

V. Related Markov Image Models

The use of neighborhoods is, of course, pervasive in the literature: they offer a geometric framework for the clustering of pixel intensities and for many types of statistical models. In particular, the Markov property is a natural way to formalize these notions. The result is a somewhat bewildering array of Markov-type image models and it seems worthwhile to puase to relate these to MRF's. The process under consideration is $F = \{F_{i,j}, (i, j) \in Z_m\}$, the gray levels, or really any pixel attribute.

An early work in this direction is Abend, Harley and Kanal [1] about pattern classification. Among many novel ideas, there is the notion of a *Markov mesh* (MM) process, in which the Markovian dependence is *causal:* generally, one assumes that, for all (i, j) and f,

$$P(F_{i,j} = f_{i,j} | F_{k,l} = f_{k,l}, (k, l) \in A_{i,j}) = P(F_{i,j} = f_{i,j} | F_{k,l} = f_{k,l}, (k, l) \in B_{i,j}) \tag{5.1}$$

where $B_{i,j} \subseteq A_{i,j} \subseteq \{(k, l): k < i \text{ or } l < j\}$. A common example is $B_{i,j} = \{(i-1, j), (i-1, j-1), (i, j-1)\}$. Besag [6], Kanal [37], and Pickard [44] also discuss such "unilateral" processes, which are usually a subclass of MRF's, although the resulting (bilateral) neighborhoods can be irregular. Anyway, for MM models the emphasis is on the causal, iterative aspects, including a recursive representation for the joint probabilities. Incidentally, a Gibbs type description of rth order Markov chains is given in [1]; of course, the full Gibbs-MRF equivalence is not perceived and was not for about five years. Derin *et al.* [14] model F as an MM process and use recursive Bayes smoothing to recover F from a noisy version $F + N$; the algorithms exploit the causality to maximize the univariate poste-

rior distribution at each pixel based on the data over a strip containing it, and are very effective at low S/N ratios for some simple images.

Motivated by a paper of Lévy [41], Woods [51] defined "P-Markov" processes for the resolution of wavenumber spectra. The definition involves two spatial regions separated by a "boundary" of width P, and correspond to the past, future, and present in 1-D. Woods also considers a family of "wide-sense" Markov fields of the form

$$F_{i,j} = \sum_{(k,l) \in W_p} \theta_{k,l} F_{i-k,j-l} + U_{i,j} \tag{5.2}$$

where $W_p = \{(k, l): 0 < k^2 + l^2 \leqslant P\}$, $\theta_{k,l}$ are the MMSE coefficients for projecting $F_{i,j}$ on $\{F_{k,l}, (k, l) \in (i, j) + W_p\}$, and $\{U_{i,j}\}$ is the error, generally nonwhite. The main theoretical result is that if $\{U_{i,j}\}$ is homogeneous, Gaussian, and satisfies a few other assumptions, then F is Gaussian, P-Markov and vice-versa. In general, there are consistency problems and the P-Markov property is hard to verify. In the nearest-neighbor case, one gets a Gaussian MRF.

Other "wide-sense" Markov processes appear in Jain and Angel [35] and Stuller and Kurz [49]. The assumptions in [35] are a nearest-neighbor system, white noise, and no blur; restoration is achieved by recursively filtering the rows $\{F_{i,j}\}_{j=1}^{m}$, which form a vector-valued, second-order Markov chain, to find the optimal interpolator of each row. In [49], causality is introduced and earlier work is generalized by considering an arbitrary "scanning pattern."

The "spatial interaction models" in Chellappa and Kashyap [10], [38] satisfy (5.2) for general coefficients and W's. The model is causal if W lies in the third quadrant. The authors consider "simultaneous autoregressive" (SAR) models, wherein the noise is white, and "conditional Markov" (CM) models, wherein the "bilateral" Markov property holds (i.e., (1.1) with $\mathcal{I}_{i,j} = (i, j) + W$) in addition to (5.2), and the noise is nonwhite. Thus, the CM models are MRF's, although in [10], [38] the boundary of Z_m is periodic, and hence boundary conditions must be adjoined to (5.2). Given any (homogeneous) SAR process there exists a unique CM process with the same spectral density, although different neighborhood structure. The converse holds in the Gaussian case but is generally false (see the discussion in Besag [6]). MMSE restoration of blurred images with additive Gaussian noise is discussed in [10], the original image is SAR or CM, usually Gaussian.

Finally, Hansen and Elliott [25] and Elliott *et al.* [17] design MAP algorithms for the segmentation of remotely sensed data with high levels of additive noise. The image model is a nearest-neighbor, binary MRF. However, the autologistic form of the joint distribution is not recognized due to the lack of the Gibbs formulation. The conditional probabilities are approximated by the product of four 1-D transitions, and segmentation is performed by dynamic programming, first for each row and then for the entire images. More recent work in Elliott *et al.* [16] is along the same lines, namely MAP estimation, via dynamic programming, of very noisy but simple images; the major differences are the use of the Gibbs formulation and improvements in the algorithms. Similar work, applied to boundary finding, can be found in Cooper and Sung [11], who use a Markov boundary model and a deterministic relaxation scheme.

VI. Maximum Entropy Restoration

There are several contact points. The Gibbs distribution can be derived (directly from physical principles in statistical mechanics) by maximizing entropy: basically, it has maximal entropy among all probability measures (equilibrium states) on Ω with the same *average* energy. Thus it is no accident that, like maximum entropy (ME) methods, ours are well-suited to nonlinear problems; see [50]. Moreover, based on the success of ME restoration (along the lines suggested by Jaynes [36]) for recovering randomly pulsed objects (cf. Frieden [19]), we intend in the future to analyze such data (e.g., starfield photographs) by our methods.

We should also like to mention the interesting observation of Trussell [50] that conventional ME restoration is a special case of MAP estimation in which the prior distribution on F is

$$P(\mathrm{F} = f) = \exp\left(-\beta \sum f_{i,j} \log f_{i,j}\right) \Big/ (\text{normalizing constant}).$$

By "conventional ME," we refer to maximizing the entropy $\Sigma f_{i,j} \log f_{i,j}$ subject to $\Sigma \eta_{i,j}^2 = \text{constant}$ ($\eta_{i,j}$ is here again the noise process); see [2]. Other ME methods (e.g., [19]) do not appear to be MAP-related.

VII. Model Selection and Parameter Estimation

The quality of the restoration will clearly depend on choices made at the modeling stage, in our case about specific energy types, attribute processes, and parameters. Cross and Jain [12] use maximum likelihood estimation in the context of Besag's [6] "coding scheme," as well as standard goodness-of-fit tests, for matching realizations of autobinomial MRF's to real textures. Kashyap and Chellappa [38] introduce some new methods for parameter estimation and the choice of neighborhoods for the SAR and CM models, mostly in the Gaussian case. These are but two examples.

For uncorrupted, simple MRF's, the coding methods do finesse the problem of the partition function. However, for more complex models and for corrupted data, we feel that the coding methods are ultimately inadequate due to the complexity of the distribution of G. This view seems to be shared by other authors, although in different contexts. Of course, for MRF's, the obstacles facing conventional statistical inference due to Z have often been noted. Even for the Ising model, analytical results are rare; a famous exception is Onsager's work on the correlational structure.

At any rate, we have developed a new method [20] for estimating clique parameters from the "noisy" data, and this will be implemented in a forthcoming paper. For now, we are obliged to choose the parameters on an ad hoc basis (which is common), but hasten to add that the quality of restoration does not seem to have been adversely affected, probably due to the relative simplicity of the MRF's we actually use for the line and intensity processes; see Section XIII.

One should also address the *general* choice of π and $\mathcal{G}$. This is really quite different than parameter estimation and somewhat related to "image understanding": how does one incorporate "real-world knowledge" into the modeling process? In

image interpretation systems, various semantical and hierarchical models have been proposed (see, e.g., [26]). We have begun our study of hierarchical Gibbs models in this paper. A *general theory* of interactive, self-adjusting models that is practical and mathematically coherent may lie far ahead.

VIII. Posterior Distribution

We now turn to the posterior distribution $P(\mathrm{F} = f, \mathrm{L} = l \mid \mathrm{G} = g)$ of the original image given the "data" g. In this section we take $S = Z_m \cup D_m$, the collection of pixel and line sites, with some neighborhood system $\mathcal{G} = \{\mathcal{G}_s, s \in S\}$; an example of such a "mixed" graph was given in Section III. The configuration space is the set of all pairs $\omega = (f, l)$ where the components of f assume values among the allowable gray levels and those of l among the (coded) line states.

We assume that X is an MRF relative to $\{S, \mathcal{G}\}$ with corresponding energy function U and potentials $\{V_C\}$:

$$P(\mathrm{F} = f, \mathrm{L} = l) = e^{-U(f,l)/T}/Z$$

$$U(f, l) = \sum_C V_C(f, l).$$

For convenience, take $T = 1$

Recall that $\mathrm{G} = \phi(H(\mathrm{F})) \odot \mathrm{N}$, where N is white Gaussian noise with mean μ and variance σ^2 and is independent of X.

We emphasize that what follows is easily extended to processes N that are more general MRF's, although we still require that N be independent of X. The operation $\odot$ is assumed invertible and we will write $\mathrm{N} = \Phi(\mathrm{G}, \phi(H(\mathrm{F}))) = \{\Phi_s, s \in Z_m\}$ to indicate this inverse.

Let $\mathcal{H}_s$, $s \in Z_m$, denote the pixels which affect the blurred image $H(\mathrm{F})$ at s. For instance, for the H in (2.3), $\mathcal{H}_s$ is the 3×3 square centered at s. Observe that Φ_s, $s \in Z_m$, depends only on g_s and $\{f_t, t \in \mathcal{H}_s\}$. By the shift-invariance of H, $\mathcal{H}_{r+s} = s + \mathcal{H}_r$ where $\mathcal{H}_r \subseteq Z_m$, $s + r \in Z_m$, and $s + \mathcal{H}_r$ is understood to be intersected with Z_m, if necessary. In addition, we will assume that $\{\mathcal{H}_s\}$ is "symmetric" in that $r \in \mathcal{H}_0 \Rightarrow -r \in \mathcal{H}_0$. Then the collection $\{\mathcal{H}_s \backslash \{s\}, s \in Z_m\}$ is a neighborhood system over Z_m. Let $\mathcal{H}^2$ denote the second-order system, i.e.,

$$\mathcal{H}_s^2 = \bigcup_{r \in \mathcal{H}_s} \mathcal{H}_r, \quad s \in Z_m.$$

Then it is not hard to see that $\{\mathcal{H}_s^2 \backslash \{s\}, s \in Z_m\}$ is also a neighborhood system. Finally, set $\mathcal{G}^P = \{\mathcal{G}_s^P, s \in S\}$ where

$$\mathcal{G}_s^P = \begin{cases} \mathcal{G}_s, & s \in D_m \\ \mathcal{G}_s \cup \mathcal{H}_s^2 \backslash \{s\}, & s \in Z_m. \end{cases} \tag{8.1}$$

The "P" stands for "posterior"; some thought shows that $\mathcal{G}^P$ is a neighborhood system on S.

Let $\boldsymbol{\mu} \in \mathbb{R}^M (M = N^2)$ have all components $= \mu$ and let $\|\cdot\|$ denote the usual norm in $\mathbb{R}^M$: $\|V\|^2 = \Sigma_1^M V_i^2$.

Theorem: For each g fixed, $P(\mathrm{X} = \omega \mid \mathrm{G} = g)$ is a Gibbs distribution over $\{S, \mathcal{G}^P\}$ with energy function

$$U^P(f, l) = U(f, l) + \|\boldsymbol{\mu} - \Phi(g, \phi(H(f)))\|^2/2\sigma^2. \tag{8.2}$$

Proof: Using standard results about "regular conditional expectations," we can and do assume that

$$P(\mathrm{X} = \omega \mid \mathrm{G} = g) = \frac{P(\mathrm{G} = g \mid \mathrm{X} = \omega) P(\mathrm{X} = \omega)}{P(\mathrm{G} = g)} \tag{8.3}$$

for all $\omega = (f, l)$, for each g.

Since $P(\mathrm{G} = g)$ is a constant and $P(\mathrm{X} = \omega) = e^{-U(\omega)}/Z$, the key term is

$$\begin{aligned} P(\mathrm{G} = g \mid \mathrm{X} = \omega) &= P(\phi(H(\mathrm{F})) \odot \mathrm{N} = g \mid \mathrm{F} = f, \mathrm{L} = l) \\ &= P(\mathrm{N} = \Phi(g, \phi(H(f))) \mid \mathrm{F} = f, \mathrm{L} = l) \\ &= P(\mathrm{N} = \Phi(g, \phi(H(f)))) \end{aligned}$$

(since N is independent of F and L)

$$= (2\pi\sigma^2)^{-M/2} \exp - \left(\frac{1}{2\sigma^2}\right) \|\boldsymbol{\mu} - \Phi\|^2.$$

We will write Φ for $\Phi(g, \phi(H(f)))$. Collecting constants we have, from (8.3),

$$P(\mathrm{X} = \omega \mid \mathrm{G} = g) = e^{-U^P(\omega)}/Z^P$$

for U^P as in (8.2); Z^P is the usual normalizing constant (which will depend on g). It remains to determine the neighborhood structure.

Intuitively, the line sites should have the *same* neighbors whereas the neighbors $\mathcal{G}_s$ of a pixel site $s \in Z_m$ should be augmented in accordance with the blurring mechanism.

Take $s \in D_m$. The local characteristics at s for the posterior distribution are, by (8.2),

$$\begin{aligned} P(L_s = l_s \mid L_r = l_r, r \neq s, r \in D_m, \mathrm{F} = f, \mathrm{G} = g) &= \frac{e^{-U^P(f,l)}}{\sum_{l_s} e^{-U^P(f,l)}} = \frac{e^{-U(f,l)}}{\sum_{l_s} e^{-U(f,l)}} \end{aligned}$$

where the sum extends over all possible values of L_s. Hence $\mathcal{G}_s^P = \mathcal{G}_s$.

For $s \in Z_m$, the term in (8.2) involving Φ does not cancel out. Now $\Phi(g, \phi(H(f))) = \{\Phi_s, s \in Z_m\}$ and let us denote the dependencies in Φ_s by writing $\Phi_s = \Phi_s(g_s; f_t, t \in \mathcal{H}_s)$. Then

$$\begin{aligned} P(F_s = f_s \mid F_r = f_r, r \neq s, r \in Z_m, \mathrm{L} = l, \mathrm{G} = g) &= \frac{e^{-U^P(f,l)}}{\sum_{f_s} e^{-U^P(f,l)}}; U^P(f, l) \\ &= U(f, l) + \sum_{r \in Z_m} (\Phi_r - \mu)^2/2\sigma^2. \end{aligned} \tag{8.4}$$

Decompose U^P as follows:

$$\begin{aligned} U^P(f, l) = &\sum_{C: s \in C} V_C(f, l) \\ &+ (2\sigma^2)^{-1} \sum_{r: s \in \mathcal{H}_r} (\Phi_r(g_r; f_t, t \in \mathcal{H}_r) - \mu)^2 \\ &+ \sum_{C: s \notin C} V_C(f, l) \\ &+ (2\sigma^2)^{-1} \sum_{r: s \notin \mathcal{H}_r} (\Phi_r(g_r; f_t, t \in \mathcal{H}_r) - \mu)^2. \end{aligned}$$

Since the last two terms do not involve f_s (remember that V_C only depends on the sites in C), the ratio in (8.4) depends only on the first two terms above. The first term depends only on coordinates of (f, l) for sites in $\mathcal{G}_s (s \in C \Rightarrow C \subseteq \mathcal{G}_s)$ and the second term only on sites in

$$\bigcup_{r : s \in \mathcal{H}_r} \mathcal{H}_r = \bigcup_{r \in \mathcal{H}_s} \mathcal{H}_r \doteq \mathcal{H}_s^2.$$

Hence, $\mathcal{G}_s^P = \mathcal{G}_s \cup \mathcal{H}_s^2 \backslash \{s\}$, as asserted in the theorem. □

IX. The Computational Problem

The posterior distribution $P(X = \omega | \boldsymbol{g})$ is a powerful tool for image analysis; in principle, we can construct the optimal (Bayesian) estimator for the original image, examine images sampled from $P(X = \omega | \boldsymbol{g})$, estimate parameters, design near-optimal statistical tests for the presence or absence of special objects, and so forth. But a conventional approach to any of these involves prohibitive computations. Specifically, our job here is to find the value(s) of ω which maximize the posterior distribution for a fixed $\boldsymbol{g}$, i.e., *minimize*

$$U(f, l) + \left\| \mu - \Phi(\boldsymbol{g}, \phi(H(f))) \right\|^2 / 2\sigma^2, (f, l) \in \Omega \qquad (9.1)$$

where (see Section VIII) Φ is defined by $\phi(H(f)) \odot \Phi = \boldsymbol{g}$. Even without L, the size of Ω is at least 2^{4000}, corresponding to a binary image on a small (64 X 64) lattice. Hence, the identification of even near-optimal solutions is extremely difficult for such a relatively complex function.

In Sections XI and XII we will describe our stochastic relaxation method for this kind of optimization. The same method works for sampling and for computing expectations (and hence forming likelihood ratios), as will be explained in Section XI. The algorithm is highly parallel, but our current implementation is serial: it uses a single processor. The restoration of more complex images than those in Section XIII, probably involving more levels in the hierarchy, may necessitate *some* parallel processing.

X. Stochastic Relaxation

There are many types of "relaxation," two of them being the type used in statistical physics and the type developed in image processing called "relaxation labeling" (RL), or sometimes "probabilistic relaxation." Basically, ours is of the former class, referred to here as SR, although there are some common features with RL.

The "Metropolis algorithm" (Metropolis *et al.* [42]) and others like it [7], [18] were invented to study the equilibrium properties, especially ensemble averages, time-evolution, and low-temperature behavior, of very large systems of essentially identical, interacting components, such as molecules in a gas or atoms in binary alloys.

Let Ω denote the possible configurations of the system; for example, $\omega \in \Omega$ might be the molecular positions or site configuration. If the system is in thermal equilibrium with its surroundings, then the probability (or "Boltzmann factor") of ω is given by

$$\pi(\omega) = e^{-\beta \mathcal{E}(\omega)} \Big/ \sum_{\omega} e^{-\beta \mathcal{E}(\omega)}, \qquad \omega \in \Omega$$

where $\mathcal{E}(\omega)$ is the potential energy of ω and $\beta = 1/KT$ where K is Boltzmann's constant and T is absolute temperature. We have already seen an example in the Ising model (4.7). Usually, one needs to compute ensemble averages of the form

$$\langle Y \rangle = \int_{\Omega} Y(\omega)\, d\pi(\omega) = \frac{\sum_{\omega} Y(\omega)\, e^{-\beta \mathcal{E}(\omega)}}{\sum_{\omega} e^{-\beta \mathcal{E}(\omega)}}$$

where Y is some variable of interest. This cannot be done analytically. In the usual Monte Carlo method, one restricts the sums above to a *sample* of ω's drawn uniformly from Ω. This, however, breaks down in the situation above: the exponential factor puts most of the mass of π over a very small part of Ω, and hence one tends to choose samples of very low probability. The idea in [42] is to choose the samples from π instead of uniformly and then weight the samples evenly instead of by $d\pi$. In other words, one obtains $\omega_1, \omega_2, \cdots, \omega_R$ from π and $\langle Y \rangle$ is approximated by the usual ergodic averages:

$$\langle Y \rangle \approx \frac{1}{R} \sum_{r=1}^{R} Y(\omega_r). \qquad (10.1)$$

Briefly, the sampling algorithm in [42] is as follows. Given the state of the system at "time" t, say $X(t)$, one randomly chooses another configuration η and computes the energy change $\Delta \mathcal{E} = \mathcal{E}(\eta) - \mathcal{E}(X(t))$ and the quantity

$$q = \frac{\pi(\eta)}{\pi(X(t))} = e^{-\beta \Delta \mathcal{E}}. \qquad (10.2)$$

If $q > 1$, the move to η is allowed and $X(t+1) = \eta$, whereas if $q \leq 1$, the transition is made *with probability* q. Thus we choose $0 \leq \xi \leq 1$ uniformly and set $X(t+1) = \eta$ if $\xi \leq q$ and $X(t+1) = X(t)$ if $\xi > q$. (A "parallel processing variant" of this for simulating certain binary MRF's is given by Berger and Bonomi [4].)

In binary, "single-flip" studies, $\eta = X(t)$ except at one site, whereas in "spin-exchange" [18] systems, a pair of neighboring sites is selected. In either case, the "flip" or "exchange" is made with probability $q/(1+q)$, where q is given in (10.2). In special cases, the single-flip system is equivalent to our Gibbs Sampler. The exchange algorithm in Cross and Jain [12] is motivated by work on the evolution of binary alloys. The samples generated are used for visual inspection and statistical testing, comparing the real and simulated textures. The model is an autobinomial MRF; see [6] or [12]. The algorithm is not suitable (nor intended) for restoration: for one thing, the intensity histogram is constant throughout the iteration process. This is necessarily the case with exchange systems which depend heavily on the initial configuration.

The algorithm in Hassner and Sklansky [28] is apparently a modification of one in Bortz *et al.* [7]. Another application of these ideas outside statistical mechanics appears in Hinton and Sejnowski [29], a paper about neural modeling but a spiritual cousin of ours. In particular, the parallel nature of these algorithms is emphasized.

The essence of every SR scheme is that changes $(\omega \to \eta)$ which *increase* energy, i.e., *lower* probability, are permitted.

By contrast, deterministic algorithms only allow jumps to states of lower energy and invariably get "stuck" in *local* minima. To get to samples from π, we must occasionally "backtrack."

All of these algorithms can be cast in a general theory involving Markov chains with state space Ω. See Hammersley and Handscomb [27] for a readable treatment. The goal is an irreducible, aperiodic chain with equilibrium measure π. If $\omega_1, \omega_2, \cdots, \omega_R$ is a realization of such a chain, then standard results yield (10.1), in fact at a rate $O(R^{-1/2})$ as $R \to \infty$. In this setup an auxiliary transition matrix is used to go from ω to η, and the general replacement recipe involves the same ratio $\pi(\eta)/\pi(\omega)$. The Markovian properties of the Gibbs Sampler will be described in the following sections.

Chemical annealing is a method for determining the low energy states of a material by a gradual lowering of temperature. The process is delicate: if T is lowered too rapidly and insufficient time is spent at temperatures near the freezing point, then the process may bog down in nonequilibrium states, corresponding to flaws in the material, etc. In *simulated* annealing, Kirkpatrick *et al.* [40] identify the solution of an optimal (computer) design problem with the ground state of an imaginary physical system, and then employ the Metropolis algorithm to reach "steady-state" at each of a decreasing sequence of temperatures $\{T_n\}$. This sequence, and the time spent at each temperature, is called an "annealing schedule." In [40], this is done on an ad hoc basis using guidelines developed for chemical annealing. Here, we prove the existence of annealing schedules which guarantee convergence to minimum energy states (see Section XII for formal definitions), and we identify the *rate* of decrease relative to the number of full sweeps.

Turning to RL, there are many similarities with SR, both in purpose and, at least abstractly, in method. RL was designed for the assignment of numeric or symbolic labels to objects in a visual system, such as intensity levels to pixels or geometric labels to cube edges, in order to achieve a "global interpretation" that is consistent with the context and certain "local constraints." Ideally, the process evolves by a series of *local* changes, which are intended to be simple, homogeneous, and performed in parallel. The local constraints are usually so-called "compatibility functions," which are much like statistical correlations, and often defined in reference to a graph. We refer the reader to Davis and Rosenfeld [13] for an expository treatment, to Rosenfeld *et al.* [46] for the origins, to Hummel and Zucker [30] for recent work on the logical and mathematical foundations, and to Rosenfeld and Kak [47] for applications to iterative segmentation.

But there are also fundamental differences. First, most variants of RL are rather ad hoc and heuristic. Second, and more importantly, RL is essentially a *nonstochastic* process, both in the interaction model and in the updating algorithms. (Indeed, various probabilistic analogies are often avoided as misleading; see [30], for example.) There is nothing in RL corresponding to an equilibrium measure or even a joint probability law over configurations, whereas there is no analogue in SR of the all-important, iterative updating *formulas* and corresponding sequence of "probability estimates" for various hypotheses involving pixel or object classification.

In summary, there are shared goals and shared features (locality, parallelism, etc.) but SR and RL are quite distinct, at least as practiced in the references made here.

XI. Gibbs Sampler: General Description

We return to the general notation of Section IV: $X = \{X_s, s \in S\}$ is an MRF over a graph $\{\mathcal{G}_s, s \in S\}$ with state spaces Λ_s, configuration space $\Omega = \prod_s \Lambda_s$, and Gibbs distribution $\pi(\omega) = e^{-U(\omega)/T}/Z$, $\omega \in \Omega$.

The general computational problems are

A) sample from the distribution π;
B) minimize U over Ω;
C) compute expected values.

Of course, we are most concerned with B), which corresponds to MAP estimation when π is the posterior distribution. The most basic problem is A), however, because A) together with annealing yields B) and A) together with the ergodic theorem yields C). We will state three theorems corresponding to A), B), and C) above. Theorem C is not used here and will be proven elsewhere; we state it because of its potential importance to other methods of restoration and to hypothesis testing.

Let us imagine a simple processor placed at each site s of the graph. The connectivity relation among the processors is determined by the bonds: the processor at s is connected to each processor for the sites in $\mathcal{G}_s$. In the cases of interest here (and elsewhere) the number of sites N is very large. However, the size of the neighborhoods, and thus the number of connections to a given processor, is modest, only eight in our experiments, including line, pixel and mixed bonds.

The state of the machine evolves by discrete changes and it is therefore convenient to discretize time, say $t = 1, 2, 3, \cdots$. At time t, the state of the processor at site s is a random variable $X_s(t)$ with values in Λ_s. The total configuration is $X(t) \doteq (X_{s_1}(t), X_{s_2}(t), \cdots, X_{s_N}(t))$, which evolves due to state changes of the individual processors. The starting configuration, $X(0)$, is arbitrary. At each epoch, only *one* site undergoes a (possible) change, so that $X(t-1)$ and $X(t)$ can differ in at most one coordinate. Let $n_1, n_2, \cdots$ be the sequence in which the sites are "visited" for replacement; thus, $n_t \subset S$ and $X_{s_i}(t) = X_{s_i}(t-1)$, $i \neq n_t$. Each processor is programmed to follow the same algorithm: at time t, a sample is drawn *from the local characteristics* of π for $s = n_t$ and $\omega = X(t-1)$. In other words, we choose a state $x \in \Lambda_{n_t}$ from the conditional distribution of X_{n_t} given the observed states of the neighboring sites $X_r(t-1)$, $r \in \mathcal{G}_{n_t}$. The new configuration $X(t)$ has $X_{n_t}(t) = x$ and $X_s(t) = X_s(t-1)$, $s \neq n_t$.

These are *local* computations, and *identical* in nature when π is homogeneous. Moreover, the actual calculation is *trivial* since the local characteristics are generally very simple. These conditional probabilities were discussed in Section IV and we refer the reader again to formulas (4.8) and (4.9). Notice that Z does not appear.

Given an initial configuration $X(0)$, we thus obtain a sequence $X(1), X(2), X(3), \cdots$ of configurations whose convergence properties will be described in Section XII. The limits obtained do not depend on $X(0)$. The sequence (n_t) we actually use is simply the one corresponding to a raster scan, i.e.

repeatedly visiting all the sites in some "natural" fixed order. Of course, in this case one does not actually need a processor at each site. But the theorems are valid for very general (not necessarily periodic) sequences (n_t) allowing for *asychronous* schemes in which each processor could be driven *by its own clock*. Let us briefly discuss such a parallel implementation of the Gibbs Sampler and its advantage over the serial version.

Computation is parallel in the sense that it is realized by simple and alike units operating largely independently. Units are dependent only to the extent that each must transmit its current state to its neighbors. Most importantly, the amount of time required for one complete update of the entire system is *independent of the number of sites.* In the raster version, we simply "move" a processor from site to site. Upon arriving at a site, this processor must first load the local neighborhood relations and state values, perform the replacement, and move on. The time required to refresh S grows linearly with $N = |S|$. Thus, for example, for the purposes at hand, the parallel procedure is potentially at least 10^4 times faster than the raster version we used, and which required considerable CPU time on a VAX 780. Of course, we recognize that the fully parallel version will require extremely sophisticated new hardware, although we understand that small prototypes of similar machines are underway at several places.

A more modest degree of parallelism can be simply implemented. Since the convergence theorems are independent of the details of the site replacement scheme $n_1, n_2, \cdots$ the graph associated with the MRF $\mathcal{X}$ can be divided into collections of sites with each collection assigned to an independently running (asynchronous) processor. Each such processor would execute a raster scan updating of its assigned sites. Communication requirements will be small if the division of the graph respects the natural topology of the scene, provided, of course, that the neighborhood systems are reasonably local. Such an implementation, with five or ten micro- or minicomputers, represents a straightforward application of available technology.

XII. Gibbs Sampler: Mathematical Foundations

As in Section XI, (n_t), $t = 1, 2, \cdots$, is the sequence in which the sites are visited for updating, and $X_s(t)$ denotes the state of site s after t replacement opportunities, of which only those for which $n_\tau = s$, $1 \leqslant \tau \leqslant t$, involve site s. For simplicity, we will assume a common state space $\Lambda_s \equiv \Lambda = \{0, 1, \cdots, L-1\}$, and as usual that $0 < \pi(\omega) < 1$ for all $\omega \in \Omega$ or, what is the same, that $\sup_\omega |U(\omega)| < \infty$. The initial configuration is $X(0)$.

We now investigate the statistical properties of the random process $\{X(t), t = 0, 1, 2, \cdots\}$. The evolution $X(t-1) \to X(t)$ of the system was explained in Section XI. In mathematical terms,

$$\begin{aligned} &P(X_s(t) = x_s, s \in S) \\ &\quad = \pi(X_{n_t} = x_{n_t} | X_s = x_s, s \neq n_t)\, P(X_s(t-1) \\ &\quad = x_s, s \neq n_t) \end{aligned} \tag{12.1}$$

where, of course, $\pi = e^{-U/T}/Z$ is the Gibbs measure which drives the process. Our first result states that the distribution of $X(t)$ converges to π as $t \to \infty$ regardless of $X(0)$. The only assumption is that we continue to visit every site, obviously a necessary condition for convergence.

Theorem A (Relaxation): Assume that for each $s \in S$, the sequence $\{n_t, t \geqslant 1\}$ contains s infinitely often. Then for every starting configuration $\eta \in \Omega$ and every $\omega \in \Omega$,

$$\lim_{t \to \infty} P(X(t) = \omega | X(0) = \eta) = \pi(\omega). \tag{12.2}$$

The proof appears in the Appendix, along with that of Theorem B. Like the Metropolis algorithm, the Gibbs Sampler produces a Markov chain $\{X(t), t = 0, 1, 2, \cdots\}$ with π as equilibrium distribution. The only complication is that the transition probabilities associated with the Gibbs Sampler are nonstationary, and their matrix representations do not commute. This precludes the usual algebraic treatment. These issues are discussed in more detail at the beginning of the Appendix.

We now turn to annealing. Hitherto the temperature has been fixed. Theorem B is an "annealing schedule" or rate of temperature decrease which forces the system into the lowest energy states. The necessary programming modification in the relaxation process is trivial, and the *local* nature of the calculations is preserved.

Let us indicate the dependence of π on T by writing π_T, and let $T(t)$ denote the temperature at stage t. The annealing procedure generates a different process $\{X(t), t = 1, 2, \cdots\}$ such that

$$\begin{aligned} &P(X_s(t) = x_s, s \in S) \\ &\quad = \pi_{T(t)}(X_{n_t} = x_{n_t} | X_s = x_s, s \neq n_t) \\ &\qquad \cdot P(X_s(t-1) = x_s, s \neq n_t). \end{aligned} \tag{12.3}$$

Let

$$\Omega_0 = \{\omega \in \Omega : U(\omega) = \min_\eta U(\eta)\}, \tag{12.4}$$

and let π_0 be the uniform distribution on Ω_0. Finally, define

$$\begin{aligned} U^* &= \max_\omega U(\omega), \\ U_* &= \min_\omega U(\omega), \\ \Delta &= U^* - U_*. \end{aligned} \tag{12.5}$$

Theorem B (Annealing): Assume that there exists an integer $\tau \geqslant N$ such that for every $t = 0, 1, 2, \cdots$ we have

$$S \subseteq \{n_{t+1}, n_{t+2}, \cdots, n_{t+\tau}\}.$$

Let $T(t)$ be any decreasing sequence of temperatures for which

a) $T(t) \to 0$ as $t \to \infty$;
b) $T(t) \geqslant N\Delta/\log t$
for all $t \geqslant t_0$ for some integer $t_0 \geqslant 2$.

Then for any starting configuration $\eta \in \Omega$ and for every $\omega \in \Omega$,

$$\lim_{t \to \infty} P(X(t) = \omega | X(0) = \eta) = \pi_0(\omega). \tag{12.6}$$

The first condition is that the individual "clocks" do not slow to an arbitrarily low frequency as the system evolves, and imposes no limitations in practice. For raster replacement,

$\tau = N$. The major practical weakness is b); we cannot truly follow the "schedule" $N\Delta/\log t$. For example, with $N = 20{,}000$ and $\Delta = 1$, it would take $e^{40{,}000}$ site visits to reach $T = 0.5$. We single out this temperature because we have obtained good results by making T decrease from approximately $T = 4$ to $T = 0.5$ over 300-1000 sweeps ($= 300N$ - $1000N$ replacements), using a schedule of the form $C/\log(1 + k)$, where k is the number of full sweeps. (Notice that the condition in b) is then satisfied provided C is sufficiently large.) Apparently, the bound in b) is far from optimal, at least as concerns the constant $N\Delta$. (In fact, the proof of Theorem B does establish something stronger, namely that Δ can be taken as the largest absolute difference in energies associated with pairs ω and ω^* which differ at only one coordinate. But this improvement still leaves $N\Delta$ too large for actual practice.) On the other hand, the logarithmic rate is not too surprising in view of the widespread experience of chemists that T must be lowered very slowly, particularly near the freezing point. Otherwise one encounters undesirable physical embodiments of *local* energy minima.

Concerning ergodicity, in statistical physics one attempts to predict the observable quantities of a system in equilibrium; these are the "time averages" of functions on Ω. Under the "ergodic hypothesis," one assuumes that (10.1) is in force, so that time averages approach the corresponding "phase averages" or expected values. The analog for our system is the assertion that, in some suitable sense,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} Y(X(t)) = \int_{\Omega} Y(\omega)\, d\pi(\omega). \tag{12.7}$$

(Here again T is fixed.) As we have already stated, a direct calculation of the righthand side of (12.7), namely,

$$\sum_{\omega} Y(\omega)\, e^{-U(\omega)/T} \Big/ \sum_{\omega} e^{-U(\omega)/T}$$

is impossible in general. The left-hand side of (12.7) suggests that we use the Gibbs Sampler and compute a time average of the function Y. For most physical systems, the ergodic hypothesis is just that—a *hypothesis*—which can rarely be verified in practice. Fortunately, for our system it is not too difficult to directly establish ergodicity.

Theorem C (Ergodicity): Assume that there exists a τ such that $S \subset \{n_{t+1}, \cdots, n_{t+\tau}\}$ for all t. Then for every function Y on Ω and for every starting configuration $\eta \in \Omega$, (12.7) *holds with probability one.*

XIII. Experimental Results

There are three groups of pictures. Each contains an original image, several degraded versions, and the corresponding restorations, usually at two stages of the annealing process to illustrate its evolution. The degradations are formed from combinations of

i) ϕ absent or $\phi(x) = \sqrt{x}$;
ii) multiplicative or additive noise;
iii) signal-to-noise levels.

The signal-to-noise ratios are all very low. For blurring, we always took the convolution H in (2.3). The restorations are all MAP estimates generated by the serial Gibbs Sampler with annealing schedule

$$T(k) = \frac{C}{\log(1 + k)}, \qquad 1 \leqslant k \leqslant K$$

where $T(k)$ is the temperature during the kth *iteration* (= full sweep of S), so that K is the total number of iterations. In each case, $C = 3.0$ or $C = 4.0$. No pre- or postfiltering, nor anything else was done. The models for the intensity and line processes were kept as simple as possible; indeed, only cliques of size two appear in the intensity model.

Group 1: The original image [Fig. 2(a)] is a sample of an MRF on Z_{128} with $L = 5$ intensities and the eight-neighbor system (Fig. 1, $c = 2$). The potentials $V_C = 0$ unless $C = \{r, s\}$, in which case

$$V_C(f) = \begin{cases} \frac{1}{3}, & f_s = f_r \\ -\frac{1}{3}, & f_s \neq f_r. \end{cases}$$

Two hundred iterations (at $T \equiv 1$) were made to generate Fig. 2(a).

The first degraded version is Fig. 2(b), which is simply Fig. 2(a) plus Gaussian noise with $\sigma = 1.5$ *relative to* gray levels f, $1 \leqslant f \leqslant 5$. Fig. 2(c) is the restoration of Fig. 2(b) with $K = 25$ iterations only, i.e., early in the annealing process. In Fig. 2(d), $K = 300$.

The second degraded image [Fig. 3(b)] uses the model

$$G = H(F)^{1/2} \cdot N \tag{13.1}$$

where $\mu = 1$ and $\sigma = 0.1$, again relative to intensities $1 \leqslant f \leqslant 5$. Fig. 3(c) and 3(d) shows the restorations of Fig. 3(b) with $K = 25$ and $K = 300$, respectively.

Group 2: Fig. 4(a) is "hand-drawn." The lattice size is 64×64 and there are three gray levels. Gaussian noise ($\mu = 0$, $\sigma = 0.7$) was added to produce Fig. 4(b). We tried two types of restoration on Fig. 4(b). First, we used the "blob process" which generated Fig. 2(a) for the F-model. There was no line process and $K = 1000$. Obviously these are flaws; see Fig. 4(c).

A line process L was then adjoined to F for the original image model, and the corresponding restoration after 1000 iterations is shown in Fig. 4(d). L itself was described in Case 2 of Section III and the neighborhood system for (F, L) on $Z_{64} \cup D_{64}$ was discussed in Case 3 of Section III. The (prior) distribution on X = (F, L) was as follows. The range of F is $\{0, 1, 2\}$ ($L = 3$ intensities). The energy $U(\boldsymbol{f}, \boldsymbol{l})$ consists of two terms, say $U(\boldsymbol{f}|\boldsymbol{l}) + U(\boldsymbol{l})$. To understand the interaction term $U(\boldsymbol{f}|\boldsymbol{l})$, let d denote a line site, say between pixels r and s. If $L_d = 1$, i.e., an edge element is "present" at d, then the bond between s and r is "broken" and we set $V_{\{r,s\}}(f_r, f_s) = 0$ regardless of f_r, f_s; otherwise ($L_d = 0$) $V_{\{r,s\}}$ is as before except that $\pm\frac{1}{3}$ are replaced by ± 1. As for $U(\boldsymbol{l})$, only cliques of size four are nonzero, of which there are six distinct types up to rotations. These are shown in Fig. 5(a) with their associated energy values.

Then we corrupted the hand-drawn figure using (13.1) with the same noise parameters as Fig. 3(b), obtaining Fig. 6(b), which is restored in Fig. 6(c) using the same prior on (F, L) as above and with $K = 1000$ iterations.

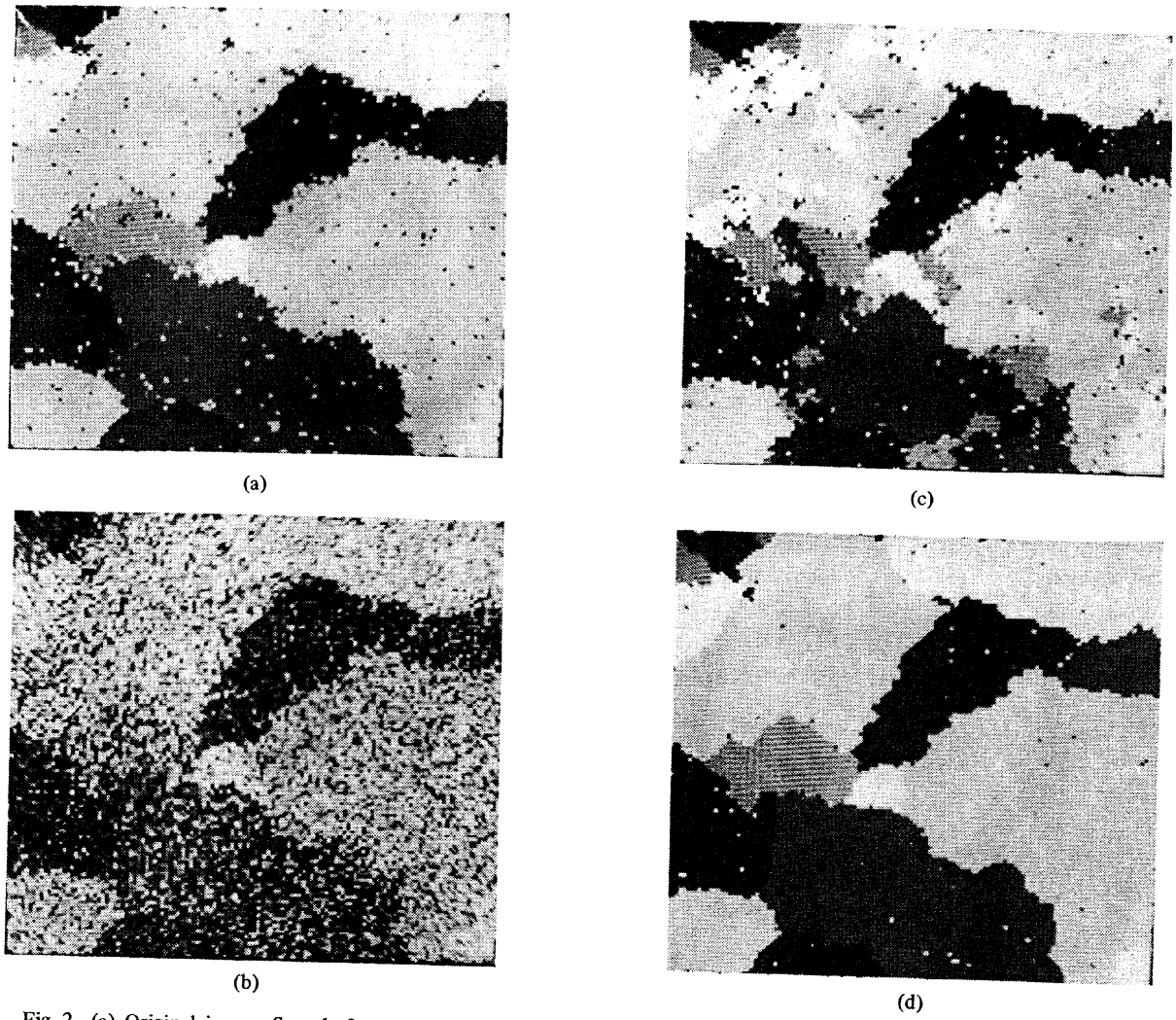


Fig. 2. (a) Original image: Sample from MRF. (b) Degraded image: Additive noise. (c) Restoration: 25 iterations. (d) Restoration: 300 iterations.

Group 3: The results in Group 2 suggest a boundary-finding algorithm for general shapes: allow the line process more directional freedom. Group 3 is an exercise in boundary finding at essentially 0 dB. Fig. 7(a) is a 64×64 segment of a roadside photograph that we obtained from the Visions Research Group at the University of Massachusetts. The levels are scaled so that the (existing) two peaks in the histogram occur at $f = 0$ and $f = 1$. We regard Fig. 7(a) as the *blurred image* $H(\mathrm{F})$. Noise is added in Fig. 7(b); the standard error is $\sigma = 0.5$ *relative to the two main gray levels* $f = 0, 1$.

Figs. 7(c) and 7(d) are "restorations" of Fig. 7(b) for $K = 100$ and $K = 1000$ iterations, respectively. The outcome of the line process is indicated by painting black any pixels to the left of or above a "broken bond." The two main regions, comprising the sign and the arrow, are perfectly circumscribed by a continuous sequence of line elements.

The model for X is more complex than the one in Group 2. There are now four possible states for each line site corresponding to "off" ($l = 0$) and three directions, shown in Fig. 5(b). The $U(f|l)$ term is the same as before in that the pixel bond between r and s is broken whenever $l_d \neq 0$. The range of F is $\{0, 1\}$ ($L = 2$).

Only cliques of size four are nonzero in $U(l)$, as before. However, there are now many combinations for $(l_{d_1}, l_{d_2}, l_{d_3}, l_{d_4})$ given such a clique $C = \{d_1, d_2, d_3, d_4\}$ of line sites, although the number is substantially reduced by assuming rotational invariance, which we do. Fig. 5(c) shows the convention we will use for the ordering and an example of the notation. The energies for the possible configurations $(l_{d_i}, 1 \leqslant i \leqslant 4)$ range from 0 to 2.70. (Remember that high energies correspond to low probability, and that the exponential exaggerates differences.) We took $V(0, 0, 0, 0) = 0$ and $V(l_{d_i}, 1 \leqslant i \leqslant 4) = 2.70$ otherwise, except when exactly two of the l_{d_i} are nonzero. Parallel segments [e.g., (1, 0, 1, 0)] receive energy 2.70; sharp turns [e.g., (0, 2, 1, 0)] and other "corner" types get 1.80; mild turns [e.g., (0, 2, 3, 0)] are 1.35; and continuations [e.g., (2, 0, 2, 0) or (0, 1, 3, 0)] are 0.90.

XIV. Concluding Remarks

We have introduced some new theoretical and processing methods for image restoration. The models and estimates are noncausal and nonlinear, and do not represent extensions into two dimensions of one-dimensional filtering and smoothing

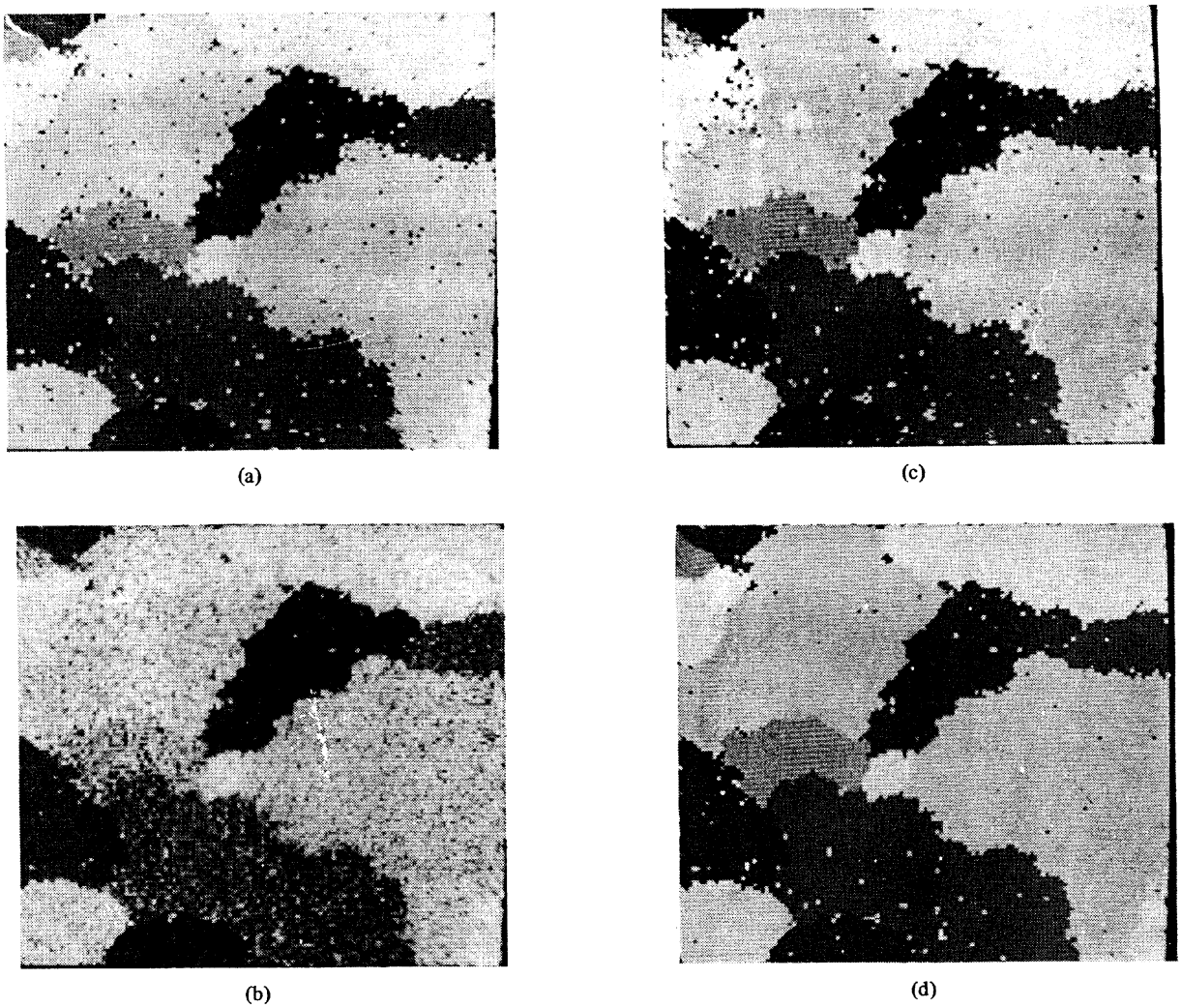


Fig. 3. (a) Original image: Sample from MRF. (b) Degraded image: Blur, nonlinear transformation, multiplicative noise. (c) Restoration: 25 iterations. (d) Restoration: 300 iterations.

algorithms. Rather, our work is largely inspired by the methods of statistical physics for investigating the time-evolution and equilibrium behavior of large, lattice-based systems.

There are, of course, *many* well-known and remarkable features of these massive, homogeneous physical systems. Among these is the evolution to minimal energy states, regardless of initial conditions. In our work posterior (Gibbs) distribution represents an *imaginary* physical system whose lowest energy states are exactly the MAP estimates of the original image given the degraded "data."

The approach is very flexible. The MRF–Gibbs class of models is tailor-made for representing the dependencies among the intensity levels of nearby pixels as well as for augmenting the usual, pixel-based process by other, unobservable attribute processes, such as our "line process," in order to bring exogenous information into the model. Moreover, the degradation model is almost unrestricted; in particular, we allow for deformations due to the image formation and recording processes. All that is required is that the posterior distribution have a "reasonable" neighborhood structure as a MRF, for in that case the computational load can be accommodated by appropriate variants (such as the Gibbs Sampler) of relaxation algorithms for dynamical systems.

Appendix

Proofs of Theorems

Background and Notation

Recall that $\Lambda = \{0, 1, 2, \cdots, L-1\}$ is the common state space, that η, η', ω, etc. denote elements of the configuration space $\Omega = \Lambda^N$, and that the sites $S = \{s_1, s_2, \cdots, s_N\}$ are visited for updating in the order $\{n_1, n_2, \cdots\} \subset S$. The resulting stochastic process is $\{X(t), t = 0, 1, 2, \cdots\}$, where $X(0)$ is the initial configuration.

For Theorem A, the transitions are governed by the Gibbs distribution $\pi(\omega) = e^{-U(\omega)/T}/Z$ in accordance with (12.1), whereas, for Theorem B (annealing), we use $\pi_{T(t)}$ (see Section XII) for the transition $X(t-1) \to X(t)$ [see (12.3)].

Let us briefly discuss the process $\{X(t), t \geq 0\}$, restricting attention to constant temperature; the annealing case is essentially the same. To begin with, $\{X(t), t \geq 0\}$ is indeed a Markov chain; this is apparent from its construction. Fix t and

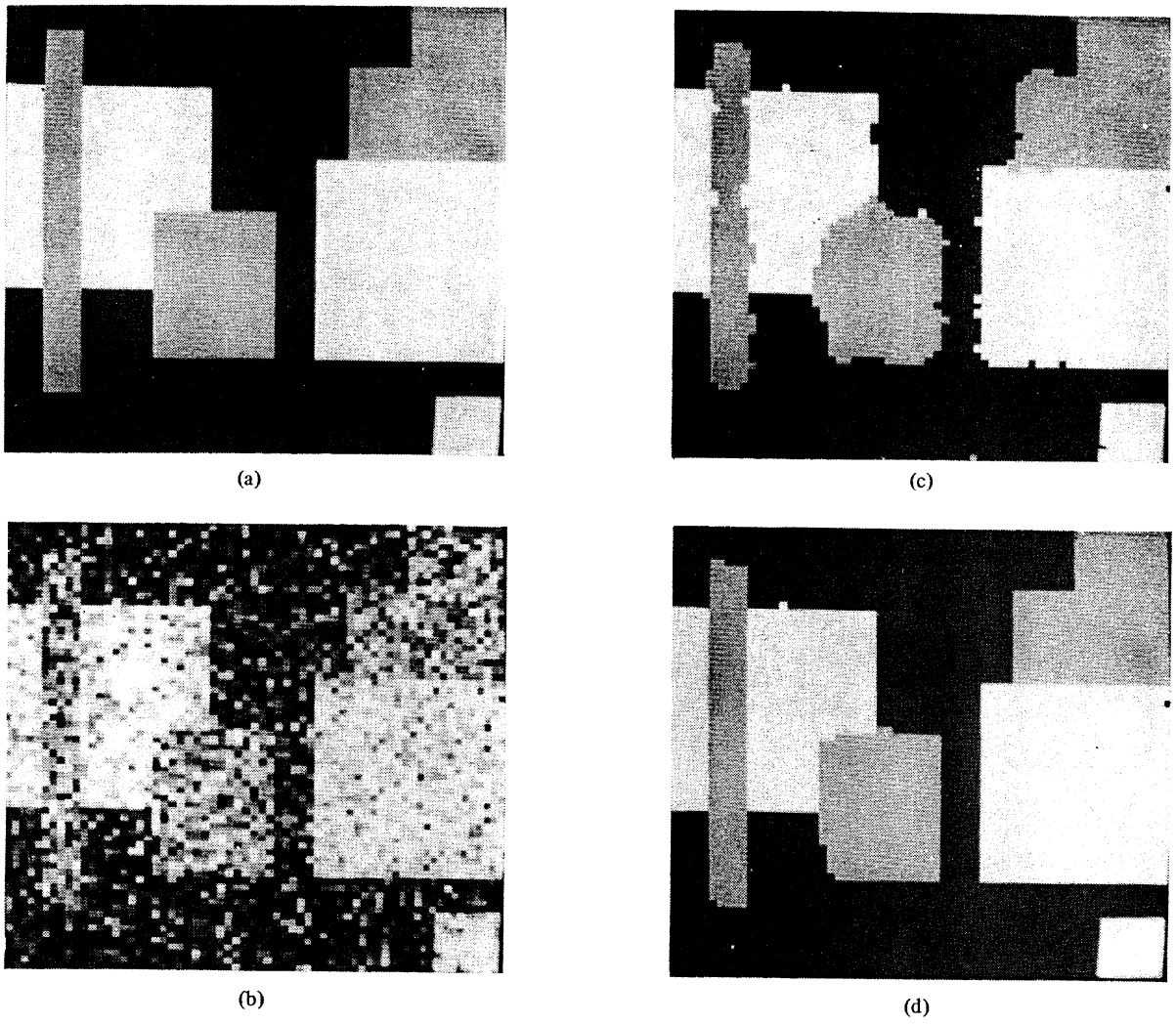


Fig. 4. (a) Original image: "Hand-drawn." (b) Degraded image: Additive noise. (c) Restoration: Without line process; 1000 iterations. (d) Restoration: Including line process; 1000 iterations.

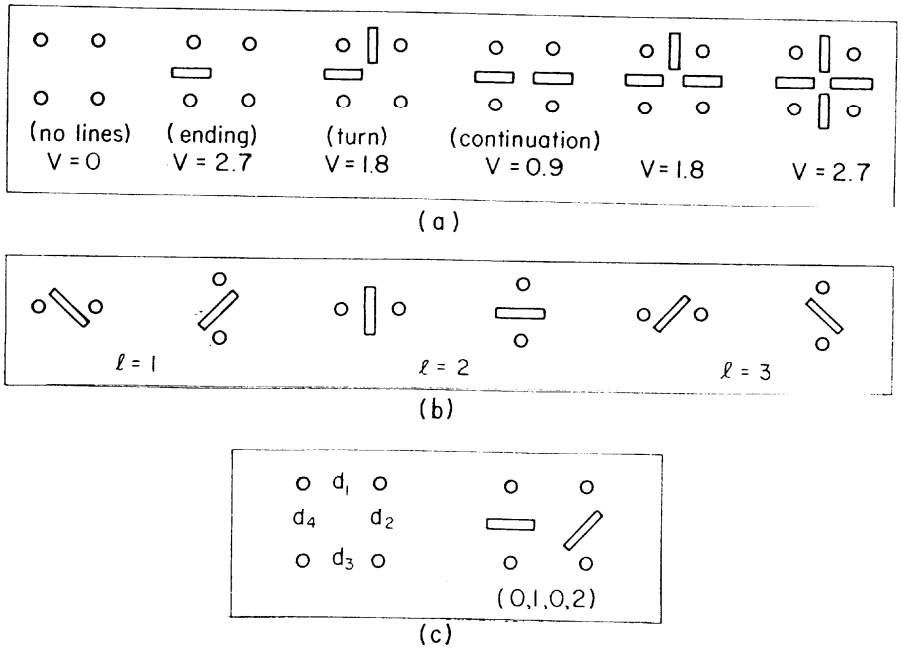
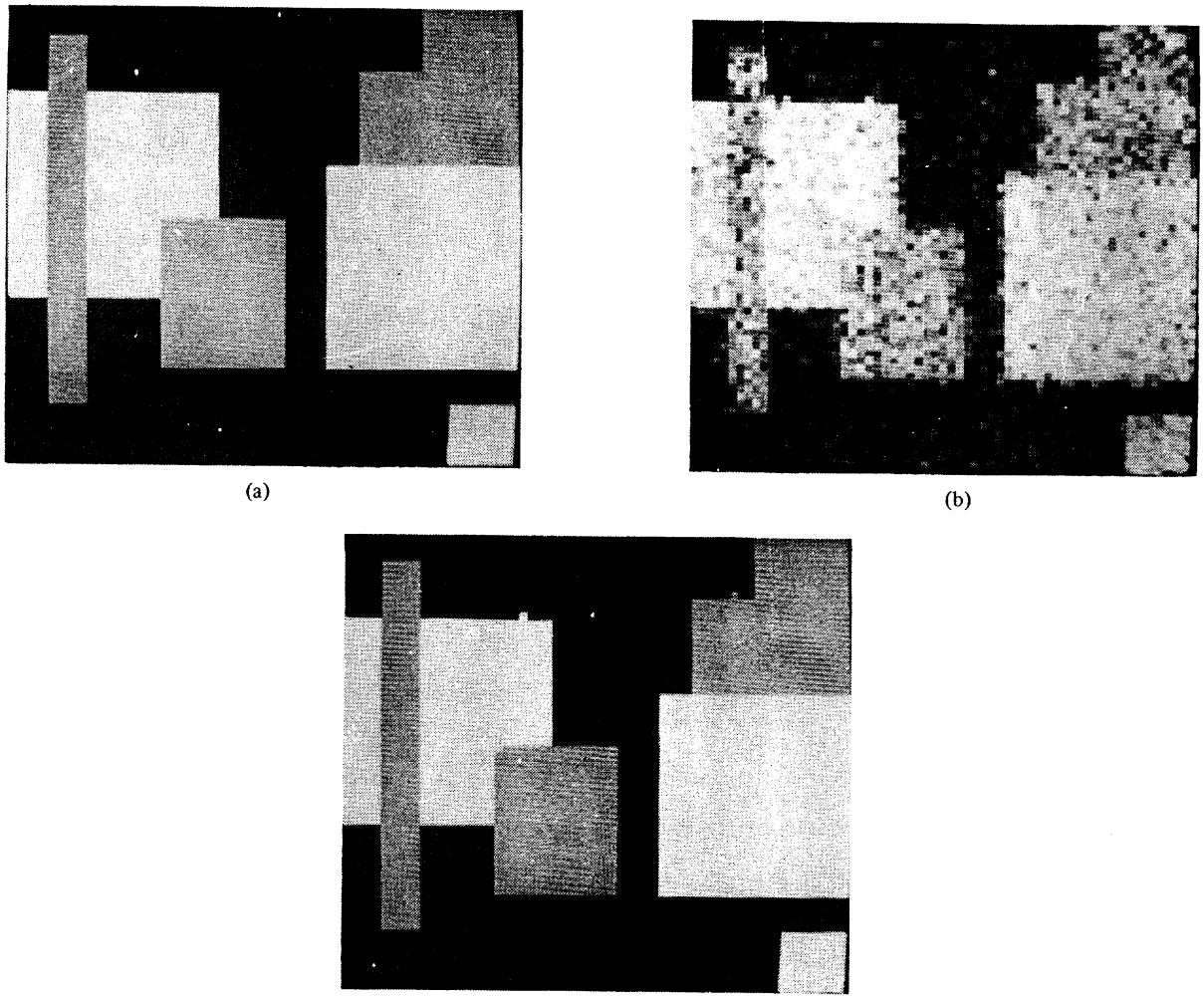


Fig. 5.



(a) (b) (c)

Fig. 6. (a) Original image: "Hand-drawn." (b) Degraded image: Blur, nonlinear transformation, multiplicative noise. (c) Restoration: including line process; 1000 iterations.

$\omega \in \Omega$. For any $x \in \Lambda$, let ω^x denote the configuration which is x at site n_t and agrees with ω elsewhere. The transition matrix *at time t* is

$$(M_t)_{\eta,\,\omega} = \begin{cases} \pi(X_{n_t} = x_{n_t} \mid X_s = x_s, s \neq n_t), \\ \qquad \text{if} \quad \eta = \omega^x \quad \text{for some} \quad x \in \Lambda \\ 0, \qquad \text{otherwise} \end{cases}$$

where $(M_t)_{\eta,\omega}$ denotes the row η, column ω entry of M_t, and $\omega = (x_{s_1}, x_{s_2}, \cdots, x_{s_N})$. In particular, the chain is *nonstationary*, although clearly *aperiodic* and *irreducible* (since $\pi(\omega) > 0\ \forall\, \omega$). Moreover, given any starting vector (distribution) μ_0, the distribution of $X(t)$ is given by the vector $\mu_0 \prod_{j=1}^{t} M_j$, i.e.,

$$P_{\mu_0}(X(t) = \omega) = \left(\mu_0 \times \prod_{j=1}^{t} M_j\right)_{\omega}$$
$$= \sum_{\eta} P(X(t) = \omega \mid X(0) = \eta)\, \mu_0(\eta).$$

Notice that π is the (necessarily) unique invariant vector, i.e., for every $t = 1, 2, \cdots$,

$$\pi(\omega) = (\pi M_t)_{\omega} = \sum_{\eta} P(X(t) = \omega \mid X(0) = \eta)\, \pi(\eta). \qquad \text{(A.1)}$$

To see this, fix t and $\omega = \{x_s\}$, and write

$$(\pi M_t)_{\omega} = \sum_{\eta} \pi(\eta)(M_t)_{\eta,\,\omega}$$
$$= \sum_{x \in \Lambda} \pi(\omega^x)(M_t)_{\omega^x,\,\omega}$$
$$= (M_t)_{\omega^{x'},\,\omega} \sum_{x \in \Lambda} \pi(\omega^x) \qquad (\text{for } \textit{any} \quad x' \in \Lambda)$$
$$= \pi(X_{n_t} = x_{n_t} \mid X_s = x_s, s \neq n_t)\, \pi(X_s = x_s, s \neq n_t)$$
$$= \pi(\omega).$$

It will be convenient to use the following, semistandard notation for transitions. For nonnegative integers $r < t$ and $\omega, \eta \in \Omega$, set

$$P(t, \omega \mid r, \eta) = P(X(t) = \omega \mid X(r) = \eta)$$

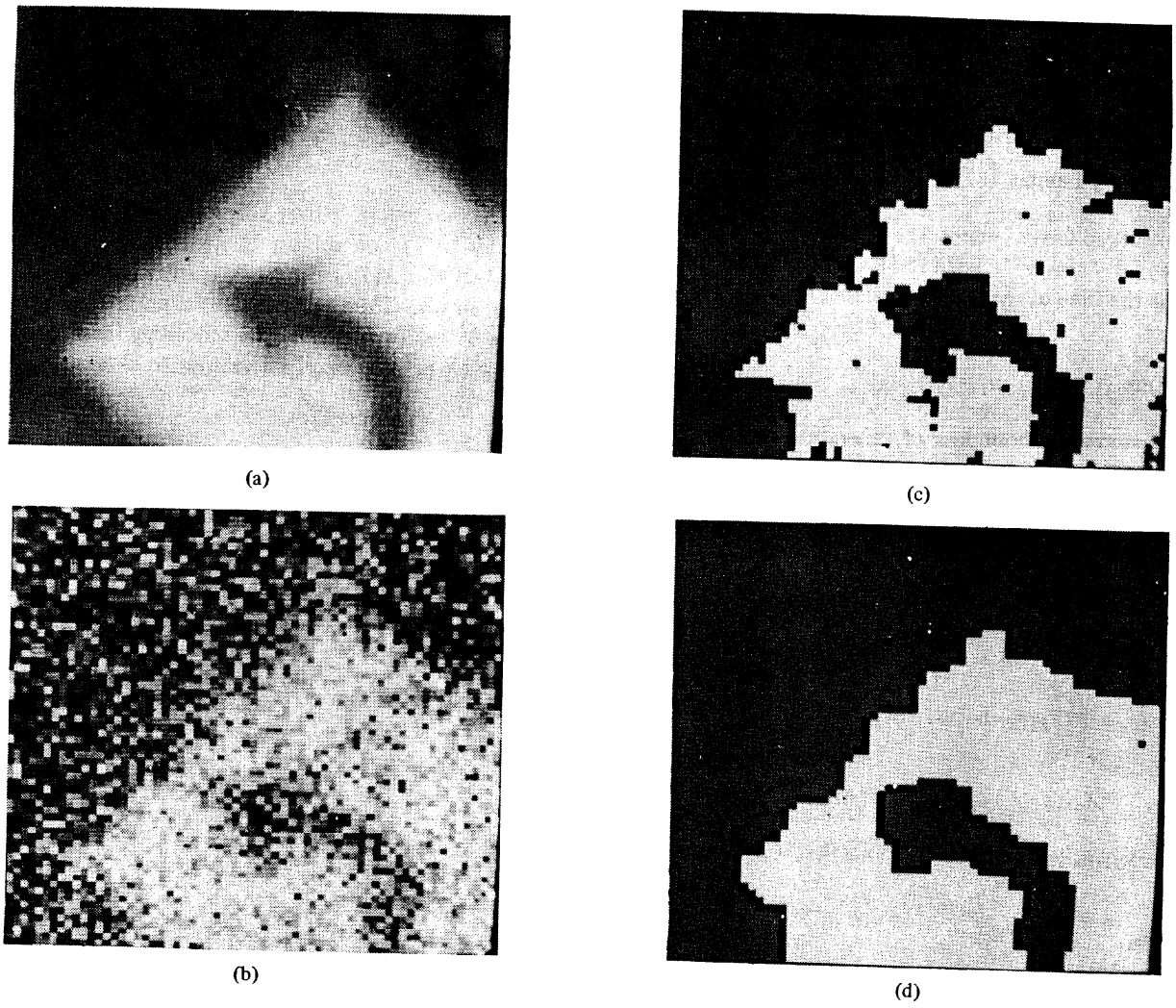


Fig. 7. (a) Blurred image (roadside scene). (b) Degraded image: Additive noise. (c) Restoration including line process; 100 iterations. (d) Restoration including line process; 1000 iterations.

and, for any *distribution* μ on Ω, set

$$P(t, \omega | r, \mu) = \sum_{\eta} P(t, \omega | r, \eta)\, \mu(\eta).$$

Finally, $||\mu - \nu||$ denotes the L^1 distance between two distributions on Ω:

$$||\mu - \nu|| = \sum_{\omega} |\mu(\omega) - \nu(\omega)|.$$

Obviously, $\mu_n \to \mu (n \to \infty)$ in distribution (i.e., $\mu_n(\omega) \to \mu(\omega)$ $\forall\, \omega$) if and only if $||\mu_n - \mu|| \to 0$, $n \to \infty$. (Remember that Ω is finite.)

Proof of Theorem A: Set $T_0 = 0$ and define $T_1 < T_2 < \cdots$ such that

$$S \subseteq \{n_{T_{k-1}+1}, n_{T_{k-1}+2}, \cdots, n_{T_k}\}, \qquad k = 1, 2, \cdots.$$

This is possible since every site is visited infinitely often. Clearly (at least) k iterations or full sweeps have been completed by "time" T_k. In particular, $kN \leqslant T_k < \infty\ \forall\, k$. Let

$$K(t) = \sup\{k : T_k < t\}.$$

Obviously $K(t) \to \infty$ at $t \to \infty$. The proof of Theorem A is based on the following lemma, which also figures in the proof of the annealing theorem.

Lemma 1: There exists a constant r, $0 \leqslant r < 1$, such that for every $t = 1, 2, \cdots$,

$$\sup_{\omega, \eta', \eta''} \left| P(X(t) = \omega | X(0) = \eta') - P(X(t) = \omega | X(0) = \eta'') \right| \leqslant r^{K(t)}.$$

Assume for now that the lemma is true. Since π is an invariant vector for the chain:

$$\begin{aligned} \overline{\lim_{t \to \infty}} \sup_{\omega, \eta} & \left| P(X(t) = \omega | X(0) = \eta) - \pi(\omega) \right| \\ &= \overline{\lim_{t \to \infty}} \sup_{\omega, \eta} \left| \sum_{\eta'} \pi(\eta') \{P(X(t) = \omega | X(0) = \eta) - P(X(t) = \omega | X(0) = \eta')\} \right| \end{aligned}$$

[by (A.1)]

$$\leqslant \overline{\lim_{t\to\infty}} \sup_{\omega,\eta',\eta''} \left| P(X(t)=\omega \,|\, X(0)=\eta') - P(X(t)=\omega \,|\, X(0)=\eta'') \right|$$

$$= 0, \text{ by Lemma 1.}$$

So it suffices to prove Lemma 1.

Proof of Lemma 1: For each $k = 1, 2, \cdots$ and $1 \leqslant i \leqslant N$, let m_i be the time of the last replacement of site s_i before $T_k + 1$, i.e.,

$$m_i = \sup \{t : t \leqslant T_k, n_t = s_i\}.$$

We can assume, without loss of generality, that $m_1 > m_2 > \cdots > m_N$; otherwise, relabel the sites. For any $\omega = (x_{s_1}, \cdots, x_{s_N})$ and ω',

$$\begin{aligned} P(X(T_k)=\omega \,|\, X(T_{k-1})=\omega') &= P(X_{s_1}(m_1)=x_{s_1}, \cdots, X_{s_N}(m_N) \\ &= x_{s_N} \,|\, X(T_{k-1})=\omega') \\ &= \prod_{j=1}^{N} P(X_{s_j}(m_j)=x_{s_j} \,|\, X_{s_{j+1}}(m_{j+1}) \\ &= x_{s_{j+1}}, \cdots, X_{s_N}(m_N)=x_{s_N}, X(T_{k-1})=\omega'). \end{aligned}$$

Let δ be the smallest probability among the local characteristics:

$$\delta = \inf_{\substack{(x_{s_1},\cdots,x_{s_N})\in\Omega \\ 1\leqslant i\leqslant N}} \pi(X_{s_i}=x_{s_i} \,|\, X_{s_j}=x_{s_j}, j\neq i).$$

Then $0 < \delta < 1$ and a little reflection shows that every term in the product above is at least δ. Hence,

$$\inf_{\substack{k=1,2,\cdots \\ \omega,\omega'}} P(X(T_k)=\omega \,|\, X(T_{k-1})=\omega') \geqslant \delta^N. \tag{A.2}$$

Consider now the inequality asserted in Lemma 1. It is trivial for $t \leqslant T_1$ since in this case $K(t) = 0$. For $t > T_1$,

$$\begin{aligned} &\sup_{\omega,\eta',\eta''} \left| P(X(t)=\omega \,|\, X(0)=\eta') - P(X(t)=\omega \,|\, X(0)=\eta'') \right| \\ &= \sup_{\omega} \{\sup_{\eta} P(X(t)=\omega \,|\, X(0)=\eta) \\ &\quad - \inf_{\eta} P(X(t)=\omega \,|\, X(0)=\eta)\} \\ &= \sup_{\omega} \Big\{ \sup_{\eta} \sum_{\omega'} P(X(t)=\omega \,|\, X(T_1) \\ &\quad = \omega') P(X(T_1)=\omega' \,|\, X(0)=\eta) \\ &\quad - \inf_{\eta} \sum_{\omega'} P(X(t)=\omega \,|\, X(T_1) \\ &\quad = \omega') P(X(T_1)=\omega' \,|\, X(0)=\eta) \Big\} \\ &\doteq \sup_{\omega} Q(t,\omega). \end{aligned}$$

Certainly, for each $\omega \in \Omega$,

$$\begin{aligned} &\sup_{\eta} \sum_{\omega'} P(X(t)=\omega \,|\, X(T_1)=\omega') P(X(T_1)=\omega' \,|\, X(0)=\eta) \\ &\qquad \leqslant \sup_{\mu} \sum_{\omega'} P(X(t)=\omega \,|\, X(T_1)=\omega') \mu(\omega') \end{aligned}$$

where the supremum is over all probability measures μ on Ω which, by (A.2), are subject to $\mu(\omega') \geqslant \delta^N \ \forall \omega'$. Suppose $\omega' \to P(X(t)=\omega \,|\, X(T_1)=\omega')$ is maximized at $\omega' = \omega^*$ (which depends on ω). Then the last supremum is attained by placing mass δ^N on each ω' and the remaining mass, namely, $1 - |\Omega| \delta^N = 1 - L^N\delta^N$, on ω^*. The value so obtained is

$$\begin{aligned} &(1 - (L^N - 1)\delta^N) P(X(t) \\ &\quad = \omega \,|\, X(T_1)=\omega^*) \\ &\quad + \delta^N \sum_{\omega' \neq \omega^*} P(X(t)=\omega \,|\, X(T_1)=\omega'). \end{aligned}$$

Similarly,

$$\begin{aligned} &\inf_{\eta} \sum_{\omega'} P(X(t)=\omega \,|\, X(T_1)=\omega') P(X(T_1)=\omega' \,|\, X(0)=\eta) \\ &\qquad \geqslant (1 - (L^N - 1)\delta^N) P(X(t) \\ &\qquad = \omega \,|\, X(T_1)=\omega_*) \\ &\qquad + \delta^N \sum_{\omega' \neq \omega_*} P(X(t)=\omega \,|\, X(T_1)=\omega_*) \end{aligned}$$

where $\omega' \to P(X(t)=\omega \,|\, X(T_1)=\omega')$ is *minimized* at ω_*. It follows immediately that

$$\begin{aligned} Q(t,\omega) &\leqslant (1 - L^N\delta^N) \{P(X(t) \\ &\quad = \omega \,|\, X(T_1)=\omega^*) - P(X(t)=\omega \,|\, X(T_1)=\omega_*)\}, \end{aligned}$$

and hence,

$$\begin{aligned} &\sup_{\omega,\eta',\eta''} \left| P(X(t)=\omega \,|\, X(0)=\eta') - P(X(t)=\omega \,|\, X(0)=\eta'') \right| \\ &\qquad \leqslant (1 - L^N\delta^N) \sup_{\omega,\eta',\eta''} \left| P(X(t) \right. \\ &\qquad = \omega \,|\, X(T_1)=\eta') - P(X(t) \\ &\qquad \left. = \omega \,|\, X(T_1)=\eta'') \right|. \end{aligned}$$

Proceeding in this way, we obtain the bound

$$\begin{aligned} &(1 - L^N\delta^N)^{K(t)} \sup_{\omega,\eta',\eta''} \left| P(X(t) \right. \\ &\qquad \left. = \omega \,|\, X(T_{K(t)})=\eta') - P(X(t)=\omega \,|\, X(T_{K(t)})=\eta'') \right| \end{aligned}$$

and the lemma now follows with $r = 1 - L^N\delta^N$. Notice that $r = 0$ corresponds to the (degenerate) case in which $\delta = L^{-1}$, i.e., all the local characteristics are uniform on Λ. Q.E.D.

Proof of Theorem B: We first state two lemmas.

Lemma 2: For every $t_0 = 0, 1, 2, \cdots$,

$$\begin{aligned} &\lim_{t\to\infty} \sup_{\omega,\eta',\eta''} \left| P(X(t) \right. \\ &\qquad \left. = \omega \,|\, X(t_0)=\eta') - P(X(t)=\omega \,|\, X(t_0)=\eta'') \right| = 0. \end{aligned}$$

Lemma 3:

$$\lim_{t_0 \to \infty} \sup_{t \geq t_0} \|P(t, \cdot \,|\, t_0, \pi_0) - \pi_0\| = 0.$$

Recall that π_0 is the uniform probability measure over the minimal energy states $\Omega_0 = \{\omega : U(\omega) = \min_\eta U(\eta)\}$.

First we show how these lemmas imply Theorem B, which states that $P(X(t) = \cdot \,|\, X(0) = \eta)$ converges to π_0 as $t \to \infty$. For any $\eta \in \Omega$,

$$\begin{aligned}
&\overline{\lim_{t \to \infty}} \|P(X(t) = \cdot \,|\, X(0) = \eta) - \pi_0\| \\
&\quad = \overline{\lim_{t_0 \to \infty}} \overline{\lim_{\substack{t \to \infty \\ t \geq t_0}}} \Big\| \sum_{\eta'} P(t, \cdot \,|\, t_0, \eta') \\
&\qquad \cdot P(t_0, \eta' \,|\, 0, \eta) - \pi_0 \Big\| \\
&\quad \leq \overline{\lim_{t_0 \to \infty}} \overline{\lim_{\substack{t \to \infty \\ t \geq t_0}}} \Big\| \sum_{\eta'} P(t, \cdot \,|\, t_0, \eta') \\
&\qquad \cdot P(t_0, \eta' \,|\, 0, \eta) - P(t, \cdot \,|\, t_0, \pi_0) \Big\| \\
&\qquad + \overline{\lim_{t_0 \to \infty}} \overline{\lim_{\substack{t \to \infty \\ t \geq t_0}}} \|P(t, \cdot \,|\, t_0, \pi_0) - \pi_0\|.
\end{aligned}$$

The last term is zero by Lemma 3. Furthermore, since $P(t_0, \cdot \,|\, 0, \eta)$ and π_0 have total mass 1, we have

$$\begin{aligned}
&\Big\| \sum_{\eta'} P(t, \cdot \,|\, t_0, \eta') P(t_0, \eta' \,|\, 0, \eta) - P(t, \cdot \,|\, t_0, \pi_0) \Big\| \\
&\quad = \sum_{\omega} \sup_{\eta''} \Big| \sum_{\eta'} (P(t, \omega \,|\, t_0, \eta') - P(t, \omega \,|\, t_0, \eta'')) \\
&\qquad \times (P(t_0, \eta' \,|\, 0, \eta) - \pi_0(\eta')) \Big| \\
&\quad \leq 2 \sum_{\omega} \sup_{\eta', \eta''} |P(t, \omega \,|\, t_0, \eta') - P(t, \omega \,|\, t_0, \eta'')|.
\end{aligned}$$

Finally, then,

$$\begin{aligned}
&\overline{\lim_{t \to \infty}} \|P(X(t) = \cdot \,|\, X(0) = \eta) - \pi_0\| \\
&\quad \leq 2 \sum_{\omega} \overline{\lim_{t_0 \to \infty}} \overline{\lim_{\substack{t \to \infty \\ t \geq t_0}}} \sup_{\eta', \eta''} |P(t, \omega \,|\, t_0, \eta') \\
&\qquad - P(t, \omega \,|\, t_0, \eta'')| \\
&\quad = 0 \quad \text{by Lemma 2.} \qquad \text{Q.E.D.}
\end{aligned}$$

Proof of Lemma 2: We follow the proof of Lemma 1. Fix $t_0 = 0, 1, \cdots$ and define $T_k = t_0 + k\tau$, $k = 0, 1, 2, \cdots$. Recall that $S \subseteq \{n_{t+1}, \cdots, n_{t+\tau}\}$ for all t by hypothesis, that $\pi_{T(t)}(\omega) = e^{-U(\omega)/T(t)}/Z$ and that U^*, U_* are the maximum and minimum of $U(\omega)$, respectively, the range being $\Delta = U^* - U_*$. Let

$$\delta(t) = \inf_{\substack{1 \leq i \leq N \\ (x_{s_1}, \cdots, x_{s_N}) \in \Omega}} \pi_{T(t)}(X_{s_i} = x_{s_i} \,|\, X_{s_j} = x_{s_j},\ j \neq i).$$

Observe that

$$\delta(t) \geq \frac{e^{-U^*/T(t)}}{L e^{-U_*/T(t)}} = \frac{1}{L} e^{-\Delta/T(t)}.$$

Now fix k for the moment and define the m_i as before:

$$m_i = \sup \{t : t \leq T_k, n_t = s_i\}, \qquad 1 \leq i \leq N.$$

We again assume that $m_1 > m_2 > \cdots > m_N$. Then

$$\begin{aligned}
&P(X(T_k) = \omega \,|\, X(T_{k-1}) = \omega') \\
&\quad = P(X_{s_1}(m_1) = x_{s_1}, \cdots, X_{s_N}(m_N) \\
&\quad = x_{s_N} \,|\, X(T_{k-1}) = \omega') \\
&\quad = \prod_{j=1}^{N} P(X_{s_j}(m_j) = x_{s_j} \,|\, X_{s_{j+1}}(m_{j+1}) \\
&\quad = x_{s_{j+1}}, \cdots, X_{s_N}(m_N) = x_{s_N}, X(T_{k-1}) = \omega') \\
&\quad \geq \prod_{j=1}^{N} \delta(m_j) \qquad \text{(using (12.3) and the definition of } \delta) \\
&\quad \geq L^{-N} \prod_{j=1}^{N} e^{-\Delta/T(m_j)} \\
&\quad \geq L^{-N} \exp - \left\{ \frac{\Delta N}{T(t_0 + k\tau)} \right\} \qquad (\text{since } m_j \leq T_k \\
&\quad = t_0 + k\tau, j = 1, 2, \cdots, N, \text{ and } T(\cdot) \text{ is decreasing}) \\
&\quad \geq L^{-N} (t_0 + k\tau)^{-1}
\end{aligned}$$

wherever $t_0 + k\tau$ is sufficiently large. In fact, for a sufficiently small constant C, we can and do assume that

$$\inf_{\omega, \omega'} P(X(T_k) = \omega \,|\, X(T_{k-1}) = \omega') \geq \frac{CL^{-N}}{t_0 + k\tau} \tag{A.3}$$

for every $t_0 = 0, 1, 2, \cdots$ and $k = 1, 2, \cdots$, bearing in mind that T_k depends on t_0.

For each $t > t_0$, define $K(t) = \sup \{k : T_k < t\}$ so that $K(t) \to \infty$ as $t \to \infty$. Fix $t > T_1$ and continue to follow the argument in Lemma 1, but using (A.3) in place of (A.2), obtaining

$$\sup_{\omega, \eta', \eta''} |P(X(t) = \omega \,|\, X(t_0) = \eta') - P(X(t) = \omega \,|\, X(t_0) = \eta'')|$$

$$\leq \prod_{k=1}^{K(t)} \left(1 - \frac{C}{t_0 + k\tau}\right).$$

Hence it will be sufficient to show that

$$\lim_{m \to \infty} \prod_{k=1}^{m} \left(1 - \frac{C}{t_0 + k\tau}\right) = 0 \tag{A.4}$$

for every t_0. However, (A.4) is a well-known consequence of the divergence of the series $\Sigma_k (t_0 + k\tau)^{-1}$ for all t_0, τ. This completes the proof of Lemma 2.

Proof of Lemma 3: The probability measures $P(t, \cdot \,|\, t_0, \pi_0)$ figure prominently in the proof, and for notational ease we prefer to write $P_{t_0, t}(\cdot)$, so that for any $t \geq t_0 > 0$ we have

$$P_{t_0, t}(\omega) = \sum_{\eta} P(X(t) = \omega \,|\, X(t_0) = \eta)\, \pi_0(\eta).$$

To begin with, we claim that for any $t > t_0 \geq 0$,

$$\|P_{t_0, t} - \pi_{T(t)}\| \leq \|P_{t_0, t-1} - \pi_{T(t)}\|. \tag{A.5}$$

Assume for convenience that $n_t = s_1$. Then

$$\begin{aligned}
&\|P_{t_0,t} - \pi_{T(t)}\| \\
&\quad = \sum_{(x_{s_1},\cdots,x_{s_N})} |\pi_{T(t)}(X_{s_1} = x_{s_1} | X_s = x_s, s \neq s_1) \\
&\qquad \cdot P_{t_0,t-1}(X_s = x_s, s \neq s_1) \\
&\qquad - \pi_{T(t)}(X_s = x_s, s \in S)| \\
&\quad = \sum_{x_{s_2},\cdots,x_{s_N}} \left\{ \sum_{x_{s_1} \in \Lambda} \pi_{T(t)}(X_{s_1} = x_{s_1} | X_s = x_s, s \neq s_1) \right. \\
&\qquad \times |P_{t_0,t-1}(X_s = x_s, s \neq s_1) \\
&\qquad \left. - \pi_{T(t)}(X_s = x_s, s \neq s_1)| \right\} \\
&\quad = \sum_{x_{s_2},\cdots,x_{s_N}} |P_{t_0,t-1}(X_s = x_s, s \neq s_1) \\
&\qquad - \pi_{T(t)}(X_s = x_s, s \neq s_1)| \\
&\quad = \sum_{x_{s_2},\cdots,x_{s_N}} \left| \sum_{x_{s_1}} \{P_{t_0,t-1}(X_s = x_s, s \in S) \right. \\
&\qquad \left. - \pi_{T(t)}(X_s = x_s, s \in S)\} \right| \\
&\quad \leq \sum_{(x_{s_1},\cdots,x_{s_N}) \in \Omega} |P_{t_0,t-1}(X_s = x_s, s \in S) \\
&\qquad - \pi_{T(t)}(X_s = x_s, s \in S)| \\
&\quad = \|P_{t_0,t-1} - \pi_{T(t)}\|.
\end{aligned}$$

Observe that $\|\pi_0 - \pi_{T(t)}\| \to 0$ as $t \to \infty$. To see this, let $|\Omega_0|$ be the size of Ω_0. Then

$$\begin{aligned}
\pi_{T(t)}(\omega) &= \frac{e^{-U(\omega)/T(t)}}{\sum_{\omega' \in \Omega_0} e^{-U(\omega')/T(t)} + \sum_{\omega' \in \Omega \setminus \Omega_0} e^{-U(\omega')/T(t)}} \\
&= \frac{e^{-(U(\omega) - U_*)/T(t)}}{|\Omega_0| + \sum_{\omega' \in \Omega \setminus \Omega_0} e^{-(U(\omega') - U_*)/T(t)}} \\
&\xrightarrow{t \to \infty} \begin{cases} 0, & \omega \notin \Omega_0 \\ \dfrac{1}{|\Omega_0|}, & \omega \in \Omega_0. \end{cases} \qquad \text{(A.6)}
\end{aligned}$$

Next, we claim that

$$\sum_{t=1}^{\infty} \|\pi_{T(t)} - \pi_{T(t+1)}\| < \infty. \qquad \text{(A.7)}$$

Since

$$\sum_{t=1}^{\infty} \|\pi_{T(t)} - \pi_{T(t+1)}\| = \sum_{\omega} \sum_{t=1}^{\infty} |\pi_{T(t)}(\omega) - \pi_{T(t+1)}(\omega)|$$

and since $\pi_{T(t)}(\omega) \to \pi_0(\omega)$ for every ω, it will be enough to show that, for every ω, $\pi_T(\omega)$ is monotone (increasing or decreasing) in T for all T sufficiently small. But this is clear from (A.6): if $\omega \notin \Omega_0$, then a little calculus shows that $\pi_T(\omega)$ is strictly increasing for $T \in (0, \epsilon)$ for some ϵ, whereas if $\omega \in \Omega_0$, then $\pi_T(\omega)$ is strictly decreasing for all $T > 0$.

Lemma 3 can now be obtained from (A.5) and (A.7) in the following way. Fix $t > t_0 \geq 0$:

$$\begin{aligned}
&\|P_{t_0,t} - \pi_0\| \\
&\quad \leq \|P_{t_0,t} - \pi_{T(t)}\| + \|\pi_{T(t)} - \pi_0\| \\
&\quad \leq \|P_{t_0,t-1} - \pi_{T(t)}\| + \|\pi_{T(t)} - \pi_0\|, \quad \text{by (A.5)} \\
&\quad \leq \|P_{t_0,t-1} - \pi_{T(t-1)}\| + \|\pi_{T(t-1)} \\
&\qquad - \pi_{T(t)}\| + \|\pi_{T(t)} - \pi_0\| \\
&\quad \leq \|P_{t_0,t-2} - \pi_{T(t-1)}\| + \|\pi_{T(t-1)} \\
&\qquad - \pi_{T(t)}\| + \|\pi_{T(t)} - \pi_0\| \\
&\quad \leq \|P_{t_0,t-2} - \pi_{T(t-2)}\| + \|\pi_{T(t-2)} \\
&\qquad - \pi_{T(t-1)}\| + \|\pi_{T(t-1)} - \pi_{T(t)}\| \\
&\qquad + \|\pi_{T(t)} - \pi_0\|.
\end{aligned}$$

Proceeding in this way,

$$\|P_{t_0,t} - \pi_0\| \leq \|P_{t_0,t_0} - \pi_{T(t_0)}\| + \sum_{k=t_0}^{t-1} \|\pi_{T(k)} - \pi_{T(k+1)}\| + \|\pi_{T(t)} - \pi_0\|.$$

Since $P_{t_0,t_0} = \pi_0$ and $\|\pi_{T(t)} - \pi_0\| \to 0$ as $t \to \infty$, we have,

$$\begin{aligned}
&\overline{\lim_{t_0 \to \infty}} \sup_{t \geq t_0} \|P_{t_0,t} - \pi_0\| \\
&\quad \leq \overline{\lim_{t_0 \to \infty}} \sup_{t > t_0} \sum_{k=t_0}^{t-1} \|\pi_{T(k)} - \pi_{T(k+1)}\| \\
&\quad = \overline{\lim_{t_0 \to \infty}} \sum_{k=t_0}^{\infty} \|\pi_{T(k)} - \pi_{T(k+1)}\| \\
&\quad = 0 \quad \text{due to (A.7).}
\end{aligned}$$

Q.E.D.

ACKNOWLEDGMENT

The authors would like to acknowledge their debt to U. Grenander for a flow of ideas; his work on pattern theory [23] prefigures much of what is here. They also thank D. E. McClure and S. Epstein for their sound advice and technical assistance, and V. Mirelli for introducing them to the practical side of image processing as well as arguing for MRF scene models.

REFERENCES

[1] K. Abend, T. J. Harley, and L. N. Kanal, "Classification of binary random patterns," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 538-544, 1965.

[2] H. C. Andrews and B. R. Hunt, *Digital Image Restoration.* Englewood Cliffs, NJ, Prentice-Hall, 1977.

[3] M. S. Bartlett, *The Statistical Analysis of Spatial Pattern.* London: Chapman and Hall, 1976.

[4] T. Berger and F. Bonomi, "Parallel updating of certain Markov random fields," preprint.

[5] J. Besag, "Nearest-neighbor systems and the auto-logistic model for binary data," *J. Royal Statist. Soc.*, series B, vol. 34, pp. 75-83, 1972.

[6] —, "Spatial interaction and the statistical analysis of lattice systems (with discussion)," *J. Royal Statist. Soc.*, series B, vol. 36, pp. 192-326, 1974.

[7] A. B. Bortz, M. H. Kalos, and J. L. Lebowitz, "A new algorithm

for Monte Carlo simulation of Ising spin systems," *J. Comp. Phys.*, vol. 17, pp. 10–18, 1975.
[8] V. Cerný, "A thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm," preprint, Inst. Phys. & Biophys., Comenius Univ., Bratislava, 1982.
[9] P. Cheeseman, "A method of computing maximum entropy probability values for expert systems," preprint.
[10] R. Chellappa and R. L. Kashyap, "Digital image restoration using spatial interaction models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 461–472, 1982.
[11] D. B. Cooper and F. P. Sung, "Multiple-window parallel adaptive boundary finding in computer vision," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-5, pp. 299–316, 1983.
[12] G. C. Cross and A. K. Jain, "Markov random field texture models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-5, pp. 25–39, 1983.
[13] L. S. Davis and A. Rosenfeld, "Cooperating processes for low-level vision: A survey," 1980.
[14] H. Derin, H. Elliott, R. Christi, and D. Geman, "Bayes smoothing algorithms for segmentation of images modelled by Markov random fields," Univ. Massachusetts Tech. Rep., Aug. 1983.
[15] R. L. Dobrushin, "The description of a random field by means of conditional probabilities and conditions of its regularity," *Theory Prob. Appl.*, vol. 13, pp. 197–224, 1968.
[16] H. Elliott, H. Derin, R. Christi, and D. Geman, "Application of the Gibbs distribution to image segmentation," Univ. Massachusetts Tech. Rep., Aug. 1983.
[17] H. Elliott, F. R. Hansen, L. Srinivasan, and M. F. Tenorio, "Application of MAP estimation techniques to image segmentation," Univ. Massachusetts Tech. Rep., 1982.
[18] P. A. Flinn, "Monte Carlo calculation of phase separation in a 2-dimensional Ising system," *J. Statist. Phys.*, vol. 10, pp. 89–97, 1974.
[19] B. R. Frieden, "Restoring with maximum likelihood and maximum entropy," *J. Opt. Soc. Amer.*, vol. 62, pp. 511–518, 1972.
[20] D. Geman and S. Geman, "Parameter estimation for some Markov random fields," Brown Univ. Tech. Rep., Aug. 1983.
[21] S. Geman, "Stochastic relaxation methods for image restoration and expert systems," in *Proc. ARO Workshop: Unsupervised Image Analysis*, Brown Univ., 1983; to appear in *Automated Image Analysis: Theory and Experiments*, D. B. Cooper, R. L. Launer, and D. E. McClure, Eds. New York: Academic, 1984.
[22] U. Grenander, *Lectures in Pattern Theory*, Vols. I–III. New York: Springer-Verlag, 1981.
[23] D. Griffeath, "Introduction to random fields," in *Denumerable Markov Chains*, Kemeny, Knapp and Snell, Eds. New York: Springer-Verlag, 1976.
[24] A. Habibi, "Two-dimensional Bayesian estimate of images," *Proc. IEEE*, vol. 60, pp. 878–883, 1972.
[25] F. R. Hansen and H. Elliott, "Image segmentation using simple Markov field models," *Comput. Graphics Image Processing*, vol. 20, pp. 101–132, 1982.
[26] A. R. Hanson and E. M. Riseman, "Segmentation of natural scenes," in *Computer Visions Systems*. New York: Academic, 1978.
[27] J. M. Hammersley and D. C. Handscomb, *Monte Carlo Methods*. London: Methuen, 1964.
[28] M. Hassner and J. Sklansky, "The use of Markov random fields as models of texture," *Comput. Graphics Image Processing*, vol. 12, pp. 357–370, 1980.
[29] G. E. Hinton and T. J. Sejnowski, "Optimal perceptual inference," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, 1983.
[30] R. A. Hummel and S. W. Zucker, "On the foundations of relaxation labeling processes," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-5, pp. 267–287, 1983.
[31] B. R. Hunt, "Bayesian methods in nonlinear digital image restoration," *IEEE Trans. Comput.*, vol. C-23, pp. 219–229, 1977.
[32] V. Isham, "An introduction to spatial point processes and Markov random fields," *Int. Statist. Rev.*, vol. 49, pp. 21–43, 1981.
[33] E. Ising, *Zeitschrift Physik*, vol. 31, p. 253, 1925.
[34] A. K. Jain, "Advances in mathematical models for image processing," *Proc. IEEE*, vol. 69, pp. 502–528, 1981.
[35] A. K. Jain and E. Angel, "Image restoration, modeling and reduction of dimensionality," *IEEE Trans. Comput.*, vol. C-23, pp. 470–476, 1974.
[36] E. T. Jaynes, "Prior probabilities," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-4, pp. 227–241, 1968.
[37] L. N. Kanal, "Markov mesh models," in *Image Modeling*. New York: Academic, 1980.
[38] R. L. Kashyap and R. Chellappa, "Estimation and choice of neighbors in spatial interaction models of images," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 60–72, 1983.
[39] R. Kinderman and J. L. Snell, *Markov Random Fields and Their Applications*. Providence, RI: Amer. Math. Soc., 1980.
[40] S. Kirkpatrick, C. D. Gellatt, Jr., and M. P. Vecchi, "Optimization by simulated annealing," IBM Thomas J. Watson Research Center, Yorktown Heights, NY, 1982.
[41] P. A. Levy, "A special problem of Brownian motion and a general theory of Gaussian random functions," in *Proc. 3rd Berkeley Symp. Math. Statist. and Prob.*, vol. 2, 1956.
[42] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equations of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, pp. 1087–1091, 1953.
[43] N. E. Nahi and T. Assefi, "Bayesian recursive image estimation," *IEEE Trans. Comput.*, vol. C-21, pp. 734–738, 1972.
[44] D. K. Pickard, "A curious binary lattice process," *J. Appl. Prob.*, vol. 14, pp. 717–731, 1977.
[45] W. H. Richardson, "Bayesian-based iterative method of image restoration," *J. Opt. Soc. Amer.*, vol. 62, pp. 55–59, 1972.
[46] A. Rosenfeld, R. A. Hummel, and S. W. Zucker, "Scene labeling by relaxation operations," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-6, pp. 420–433, 197.
[47] A. Rosenfeld and A. C. Kak, *Digital Picture Processing*, vols. 1, 2, 2nd ed. New York: Academic, 1982.
[48] F. Spitzer, "Markov random fields and Gibbs ensembles," *Amer. Math. Mon.*, vol. 78, pp. 142–154, 1971.
[49] J. A. Stuller and B. Kruz, "Two-dimensional Markov representations of sampled images," *IEEE Trans. Commun.*, vol. COM-24, pp. 1148–1152, 1976.
[50] H. J. Trussell, "The relationship between image restoration by the maximum a posteriori method and a maximum entropy method," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 114–117, 1980.
[51] J. W. Woods, "Two-dimensional discrete Markovian fields," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 232–240, 1972.

Stuart Geman received the B.A. degree in physics from the University of Michigan in 1971, the M.S. degree in physiology from Dartmouth College in 1973, and the Ph.D. degree in applied mathematics from the Massachusetts Institute of Technology in 1977.

Since 1977 he has been a member of the Division of Applied Mathematics at Brown University, Providence, RI, where he is currently an Associate Professor. His research interests include statistical inference, parallel computing, image processing, and stochastic processes.

Dr. Geman is an Associate Editor of *The Annals of Statistics* and is a recipient of the Presidential Young Investigator Award.

Donald Geman, for a photograph and biography, see this issue, p. 720.