November 1992 LIDS-P-2167

Modelling and Estimation for Random Fields

Sanjoy K. Mitter
Department of Electrical Engineering and Computer Science
and
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge MA 02139-4307 USA

1. Introduction

Filtering of random processes is now a well developed subject. To describe the situation, consider a Markov process $x(t)$ whose evolution is described by a stochastic differential equation

$$dx(t) = f(x(t))\,dt + \sigma(x(t))dw(t), \tag{1.1}$$

where $x(t) \in \mathbf{R^n}$, $w(t)$ is m-dimensional Brownian motion (that is $\frac{dw}{dt}$ is white Gaussian noise), $f : \mathbf{R}^n \to \mathbf{R}^n$, the drift, is a smooth bounded function and $\sigma : \mathbf{R}^n \to \mathbf{R}^n \times \mathbf{R}^m$, the diffusion matrix is also smooth and bounded such that the matrix function $\sigma(\cdot)\sigma^T(\cdot)$ is invertible. We assume that we cannot observe $x(\cdot)$ directly but we observe a non-linear function of $x(\cdot)$ in the presence of white noise, that is, we observe

$$dy(t) = h(x(t))dt + dv(t), \text{ where } h : \mathbf{R}^n \to \mathbf{R}^p \tag{1.2}$$

in smooth and $E[\int_0^T h(x(t))^2 dt] < \infty$, and $v(t)$ is also standard p-dimensional Brownian motion which is independent of $w(t)$. The filtering problem is the following:

By observing $y(\cdot)$ on the interval $[0, T]$, we are required to estimate $x(T)$, and this estimate is to be built recursively, in the sense that $x(T)$ is estimated on the basis of past data, where $T > 0$ is arbitrary, such that the estimate on the interval $[0, T + s]$ is computed on the basis of the estimate of $x(T)$ and the new observation on the interval $[T, T + s]$. What makes this possible is the assumption that $x(t)$ is a Markov process and hence the conditional distribution of $x(t)$ given the past is the same as the conditional distribution of $x(t)$ given the immediate past. Thus $x(t)$ has probabilistically a local character and this is exploited in the recursive computation of the estimate of $x(t)$. Now, whatever our definition of estimate is, it can be computed by computing the conditional density (assumed to exist) $p(t, x|\pi_t y)$ (where $\pi_t y$ denotes the past of $y(\cdot)$) and describing its evolution. Thus the *filter* could be considered as a mapping from $\pi_t y \to p(t, x|\pi_t y)$. It turns out that $p(t, x|\pi_t y)$ can be written in terms of what may be called the unnormalized conditional density

$$p(t, x|\pi_t y) = \frac{\rho(t, x|\pi_t y)}{\int_{\mathbf{R}^n} \rho(t, x|\pi_t y)dx} \tag{1.3}$$

and $\rho(t, x|\pi_t y)$ satisfies a stochastic partial differential equation

$$d\rho(t, x|\pi_t y) = \mathcal{L}_0^* \rho(t, x|\pi_t y)\, dt + \mathcal{L}_1 \rho(t, x|\pi_t y)\, dy(t) \tag{1.4}$$

where

$$\mathcal{L}_0 = \frac{1}{2}\sum_{i,j} a_{ij}(x)\frac{\partial^2}{\partial x_i \partial x_j} + \sum_i f_i(x)\frac{\partial}{\partial x_i} \tag{1.5}$$

($a_{ij}(\cdot)$ is the $i,j^{\underline{th}}$ element of the matrix-value function $\sigma^T(x)\sigma(x)$ and f_i is the $i^{\underline{th}}$ component of the vector-valued function $f(x)$),

$\mathcal{L}_0^*$ is the formal adjoint of $\mathcal{L}_0$ (1.6)

$\mathcal{L}_1$ is the operator which is multiplication by the function $h(x)$. (1.7)

The special case of this situation is the Gauss-Markov case where

$$\begin{cases} f(x(t)) = Ax(t), & A = n \times n \text{ matrix} \\ \sigma(x(t)) = B, & \text{an } n \times m \text{ matrix} \\ h(x(t)) = Cx(t), & C = p \times n \text{ matrix} \end{cases} \tag{1.8}$$

In this case, $p(t, x|\pi_t y)$ is conditionally Gaussian and hence can be completely described by describing the evolution of the conditional mean $\hat{x}(t) = E[x(t)|\pi_t y]$ (and the conditional covariance $\Sigma(t)$. The evolution of $\hat{x}(t)$ is described by

$$d\hat{x}(t) = A\hat{x}(t)dt + K(t)[dy(t) - C\hat{x}(t)dt] \tag{1.9}$$

$K(t)$ is characterized by the covariance of the error $e(t) = x(t) - \hat{x}(t)$ and is independent of $y(\cdot)$. This is the celebrated Kalman-Bucy filter.

It is worth remarking that the coupled $(\hat{x}(t), y(t))$ process is a Markov process.

Much of the theory of Kalman-Bucy Filtering can be carried over to systems described by linear partial differential equations, provided we are willing to deal with the intricacies of the Wiener process with values in infinite-dimensional spaces. This is done by writing the partial differential equation as an abstract evolution equation in an appropriate Hilbert space.

Consider the acoustic wave equation

$$\begin{cases} u_{tt} = c_0^2 \Delta u(x,t) \\ u(x,0) = f(x), \quad u_t(x,0) = g(x) \end{cases} \tag{1.10}$$

Here Δ is the 3-dimensional Laplacian and c_0 is the velocity of propagation of pressure waves.

To formulate this as an abstract evolution, consider the operator $H_0 = -c_0^2\Delta$ on $L^2(\mathbf{R}^3)$ and $B_0 = \sqrt{H_0}$. Denote by $\overline{D(B_0)}$ the closure of $D(B_0)$ in the norm $||B_0 u||_2$, the L^2-norm.

Let $\mathcal{H}_0 = \overline{D(B_0)} \oplus L^2(\mathbf{R}^3)$, with the norm

$$||(u,v)||^2 = ||B_0 u||_2^2 + ||v||_2^2$$

and define

$$A_0 = i\begin{pmatrix} 0 & I \\ -B_0^2 & 0 \end{pmatrix}, D(A_0) = D(B_0^2) \oplus D(B_0)$$

where $D(B_0^2) = \{u \in \overline{D(B_0)} | B_0 u \in D(B_0)\}$ (both B_0 and its extension to $\overline{D(B_0)}$ are denoted by B_0). A_0 is a self-adjoint operator on $D(A_0)$ and the wave equation can be written as

$$\begin{cases} \dot{\phi}(t) = -iA_0\phi(t), \\ \phi(0) = \phi_0 := (f, g) \in D(A_0). \end{cases} \tag{1.11}$$

for the $\mathcal{H}_0$-valued function $\phi(t) = (u(t), u_t(t))$. The solution is given by $\phi(t) = W_0(t)\phi_0$ where

$$W_0(t) = \begin{pmatrix} \cos B_0 t & B_0^{-1} \sin B_0 t \\ -B_0 \sin B_0 t & \cos B_0 t \end{pmatrix}$$

where the matrix entries are defined using functional calculus.

We wish to describe a corresponding problem for random fields, that is; a process which is indexed not by time (a totally ordered set) but by a set (e.g. $\mathbf{R}^2$) on which there is no natural ordering. Guided by the previous development we may conjecture that we need the analogue of the Markov property. This is provided by the theory of Gibbs fields which in many situations is equivalent to so-called markov random fields.

2. Markov Random Fields on a Finite Lattice

Let $S = \{s_1, \cdots, s_N\}$ be a finite set of sites. We shall consider variable $x = (x_s)_{s \in S}$ where each $x_s \in \Sigma_s \subset \mathbf{R}$ and let $\Omega = \Pi_{s \in S} \Sigma_s$, the configuration space. We shall also have occasion to write x as $x = (x_1, \cdots, x_N)$. Let X_s denote the coordinate variables on Ω and let P be a probability measure on Ω satisfying $P(x) > 0 \ \forall \ x \in \Omega$. If $A \subset S$ then the conditional probabilities $P(X_s = x_s, s \in A | X_s = x_s, s \notin A)$ are well-defined. The one-dimensional probability distributions

$$\begin{aligned} &P(X_s = \lambda | X_r = x_r, r \neq s), \ s \in S, \ x \in \Omega \\ &:= P_s(\lambda | X_{(s)}) \text{ where } \lambda = x_s \text{ and } x_{(s)} = (x_r)_{r \neq s} \end{aligned}$$

then determine the distribution of X. We shall see a generalization of this idea when Card (S) is not finite in the next section.

Let $\mathcal{P}(S)$ denote the set of subsets of S. A *neighbourhood system* is a collection $\mathcal{N} = (\mathcal{N}_s)_{s \in S}$ where $\mathcal{N}_s \in \mathcal{P}(S)$ and $s \notin \mathcal{N}_s$ and $s \in \mathcal{N}_t \Leftrightarrow t \in \mathcal{N}_s$. The pair $(S, \mathcal{N})$ is then a graph whose vertices are the sites $s \in S$ and the edges are the pair (s, t) where $s \in \mathcal{N}_t$.

A *Markov random field* with respect to $\mathcal{N}$ is a process $(X_s)_{s \in S}$ with distribution P such that

$$P_s(x_s | x_{(s)}) = P(x_s | x_r, r \in \mathcal{N}_s) \ \forall \ s \in S, x \in \Omega.$$

A *Gibbsian random field* is a representation of a Markov random field via potentials.

A *potential* is a family $V = \{V_A : A \subset S\}$ where $V_A : \Omega \to \mathbf{R}$ such that $V_\phi = 0$ and $V_A(x) = V_A(x')$ if $x_s = x'_s$, $\forall \ s \in A$. V is said to be normalized if $V_A(x) = 0$ whenever $x_t = 0$, $t \in A$ and we assume $0 \in \Sigma_s$, $\forall \ s$.

The *energy* (Hamiltonian) associated with V is

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

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

$$P(x) = Z^{-1} e^{-H(x)}, Z = \sum_x e^{-H(x)} < \infty$$

and $V_A = 0, \forall \ A \notin \mathcal{C}$ and $H(x) = \sum_{C \in \mathcal{C}} V_C(x)$.

EXAMPLE 2.1. *2-D Ising Model*

Let $S = \{(i,j)|1 \leq i,j \leq N\}$. *Let* $\mathcal{N}$ *be the nearest neighbour system* $\{(i-1,j),(i+1,j),(i,j-1),(i,j+1)\} \cap S$

$$\Sigma_{i,j} = \{-1,+1\} \text{ and } H(x) = -\frac{h\cdot}{T}\sum_s x_s - \frac{J\cdot}{T}\sum_{<s,t>} x_s x_t$$

where $< s,t >$ *denotes nearest neighbour pair,* T *is the temperature,* h *strength of the external field and* J *is the coupling coefficient with* $J > 0$ *corresponding to the attractive case and* $J < 0$ *is the repulsive case.*

EXAMPLE 2.2. *Spin Glass*

In this case, with the same definitions as in Example 1 the Hamiltonian is given by

$$H(x) = \sum_{<s,t>} \eta_{st} x_s x_t$$

where (η_{st}) *is another random field independent of* x.

It turns out that any probability measure $P > 0$ can be expressed as a Gibbs distribution with respect to a canonical potential.

Let us introduce the following notation: For $x \in \Omega$, $A \subset S$, let

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

THEOREM 2.1. *An* $P > 0$ *is a Gibbs distribution with respect to the canonical potential*

$$V_A(x) = \sum_{B \subset A} (-1)^{|A-B|} \log P(x^B), |A-B| = \textit{Card } (A/B). \tag{2.1}$$

Moreover for any element $s \in A$

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

The representation is unique amongst normalized potentials.

The proof of this theorem follows from the Möbius Inversion Formula (see Section 3 for the Möbius Inversion Formula).

We finally come to the main theorem of this section.

THEOREM 2.2. *Let* $\mathcal{N}$ *be a neighbourhood system. Then* P *is a Gibbs distribution with respect to* $\mathcal{N}$ *iff* P *is a Markov Random field with respect to* $\mathcal{N}$, *in which case* (V_A) *in (2.2) satisfies* $V_A = 0, \forall\ A \notin \mathcal{C}$ *and* $H(x) = -\sum_{c \in \mathcal{C}} V_c(x)$.

Sketch of **Proof.** Let P have a Gibbs distribution with respect to $\mathcal{N}$ for some V, that is,

$$P(x) = Z^{-1}\mathcal{C}^{-H(x)}, \text{and } H(x) = -\sum_{C \in \mathcal{C}} V_C(x).$$

For $x \in \Omega, s \in S, \lambda \in \Sigma_s$, let $(\lambda, x_{(x)})$ denote the configuration where x_s has been replaced by λ.

Then

$$P_s(x_s|x_{(s)}) = \frac{\exp(-H_V(x))}{\sum_{\lambda \in \Sigma_x} \exp(-H_V(\lambda, x_{(s)}))}$$

$$= \frac{\exp(\sum_{A \in \mathcal{C}, s \in A} V_A(x))}{\sum_{\lambda \in \Sigma_s} \exp(\sum_{A \in \mathcal{C}, s \in A} V_A(\lambda, x_{(s)})}$$

Now $A \in \mathcal{C}$ and $s \in A$ implies that $A \subset \mathcal{N}_s + s$. Hence $P_s(x_s|x_{(s)})$ depends only on x_t, $t \in \mathcal{N}_s + s$ and hence $P_s(x_s|x_{(s)}) = P(x_s|x_r, r \in \mathcal{N}_s)$.

Now suppose that P is a Markov Random field with respect to $\mathcal{N}$ with $V = (V_A)$ the canonical potential of Theorem 2.1. The proof is completed by using the formula (2.2) and showing $V_A(x) = 0$ if $A \notin \mathcal{C}$. ∎

It is instructive to consider Markov chains as special cases of Markov Random fields and equivalently Gibbs fields. Let $\{X_n | 0 \leq n \leq N\}$ be a Markov process with state space $\Sigma, P(X_0 = \lambda) = \mu(\lambda) > 0$ and transition probabilities $P_n(\alpha, \beta) = P(X_{n+1} = \beta | X_n = \alpha) > 0 \ \forall \ \alpha, \beta \in \Sigma$. Define $\mathcal{N}_0 = \{1\}, \mathcal{N}_n = \{n-1, n+1\}, 1 \leq n \leq N-1, \mathcal{N}_N = \{N-1\}$. Then (X_n) is a Markov Random field with respect to $\mathcal{N} = (\mathcal{N}_k)_{k=0}^N$. The one-dimensional, conditional distributions are

$$P_0(x_0|x_{(0)}) = \frac{\mu(x_0)P_0(x_0, x_1)}{\sum_{\sigma \in \Sigma} \mu(\sigma)P_0(\sigma; x_1)}$$

$$P_n(x_n|x_{(n)}) = \frac{P_{n-1}(x_{n-1}, x_n)P_n(x_n, x_{n+1})}{\sum_{\sigma \in \Sigma} P_{n-1}(x_{n-1}, \sigma)P_n(\sigma, x_{n+1})}, 1 \leq n \leq N-1$$

$$P_N(x_N|x_{(N)}) = P_{N-1}(x_{N-1}, x_N)$$

Therefore, the one-sided Markov property implies the two-sided Markov property. Now consider the Markov process $\{X_n | 0 \leq n \leq N\}$ which has the two-sided Markov property. Then it is a Markov Random field with respect to the neighbourhood system $\mathcal{N} = (\mathcal{N}_n)$ where $\mathcal{N}_n = \{n-1, n+1\}$. It then has an associated canonical potential $V_A(x)$ for the clique $A = \{n-1, n\}$ given by

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

2.1. BAYESIAN ESTIMATION WITH MARKOV RANDOM FIELD PRIOR MODELS.

The problem of interest here is the estimation of a process X which is a Markov Random Field with respect to a neighbourhood system $\mathcal{N} = (\mathcal{N}_s)$ from observations Y which is taken to be a local function, possibly non-linear, of X and corrupted by noise. If P is the Gibbs distribution corresponding to the Markov field X, then the estimation problem corresponds to computing the conditional distribution $P(x|y)$ which can be computed from the Bayes formula

$$P(x|y) = \frac{P(x)P(y|x)}{P(y)}.$$

Under the assumption that Y is a local function of X corrupted by noise, it turns out that the conditional distribution is again a Gibbs distribution and hence a Markov Random Field with a different neighbourhood structure $\mathcal{N}'$.

To illustrate how this works, let X correspond to the Ising model on $S \subset Z^2$ with states $\{-1, +1\}$ with free boundary conditions and no external field. Hence

$$P(X = x) = Z_T^{-1} \exp(\frac{1}{T} \sum_{<s,t>} x_s x_t).$$

The noisy observation process corresponds to a binary symmetric channel given by $Y_s = X_s W_s, s \in S$, where W and X are independent, (W_s) is i.i.d with

$$P(W_s = -1) = \varepsilon = 1 - P(W_s = +1)$$

Then the conditional distribution is given by

$$P(x|y) = Z_{T,y}^{-1} \exp\{\frac{1}{T} \sum_{<s,t>} x_s x_t + \frac{1}{2} \log(\frac{1-\varepsilon}{\varepsilon}) \sum_s x_s y_s\}$$

If we denote by $\hat{x}$ an estimate of x, then there are several choices of loss functions for choosing the estimate.

A Bayes estimate for the loss function

$$L(x, \hat{x}) = \sum_{s \in S} 1_{\hat{x}_s = x_s},$$

is given by

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

Another possibility of an estimate is obtained by

$$\begin{aligned} \hat{x} &= \arg\max_x P(X = x|Y = y) \\ &= \arg\min_x \{-\frac{1}{T} \sum_{<s,t>} x_s x_t - \frac{1}{2} \log(\frac{1-\varepsilon}{\varepsilon}) \sum_s x_s y_s\} \end{aligned}$$

This is the so-called maximum a posteriori probability estimate (MAP). The interest in global optimization algorithms which we discuss in a later section arises when we wish to compute MAP estimates.

3. Gibbs Fields and Gibbs Measures

In Section 2 we have described Markov Random fields on a finite set of sites. To describe such fields on a possibly infinite set of sites one starts with Gibbs fields and Gibbs measures and then deduce their Markovian properties.

We follow here Dobrushin [1] to describe the basic theory of Gibbs measures and its relation to Markovian properties of these measures.

Let $G = (S, E)$ be a denumerable graph consisting of vertices (sites) $S = (\cdots i, j, k, \cdots)$ and edges E. We say that a pair (i, j) is a neighbour if (i, j) is an edge. We shall usually work in the case where $G = Z^d$, $d > 1$, the d-dimensional lattice. We shall equip Z^d with the distance between $i = (i_1, \cdots, i_d)$ and $j = (j_1, \cdots, j_d)$ given by $\rho(i, j) = \sum_{k=1}^{d} |i_k - j_k|$. i and j are said to be neighbours if $\rho(i, j) = 1$.

At each vertex (site) $i \in S$, we consider a (random) variable $X(i)$ taking values in the space Σ. This then defines a mapping $X : S \to \Sigma : i \to X(i)$. Σ will in general be Polish space (for example, a finite set, $\mathbf{R}, \mathbf{R}^n$, the sphere S^1 etc.). We denote by $\Omega = \Pi_{i \in S} \Sigma_i$, where Σ_i is a copy of Σ.

An element of Ω is called a configuration.

Consider the measurable space $(\Sigma, \mathcal{B}_\Sigma)$ where $\mathcal{B}_\Sigma$ is the Borel σ-field of Σ. We shall now use a number of properties of such spaces (see Parthasarathy [2])

a) $\mathcal{B}_\Sigma$ has a denumerable sub-family $\mathcal{D}$ such that $\mathcal{B}_\Sigma$ is the smallest σ-algebra of subsets of Σ containing $\mathcal{D}$.

b) Let $\Sigma' \subseteq \Sigma$. Then $\mathcal{B}_{\Sigma'} = \{E \cap \Sigma' | E \in \mathcal{B}_\Sigma\}$ In particular if Σ is a Borel set in Σ, then $\mathcal{B}'_\Sigma$ is precisely the class of all subsets of Σ' which are Borel sets in Σ.

c) Let $\Sigma_1, \Sigma_2, \cdots$ be separable metric spaces and Σ their cartesian product. Then the Borel space $(\Sigma, \mathcal{B}_\Sigma)$ is the cartesian product of the Borel spaces $(\Sigma_n, \mathcal{B}_{\Sigma_n}), n = 1, 2, \cdots$

Using the above properties we consider the measurable space $(\Omega, \mathcal{B}_\Omega)$. A random field is a probability measure μ on $(\Omega, \mathcal{B}_\Omega)$.

If $A \subset S$, then $\Omega_A := \Pi_{i \in A} \Sigma_i$ and μ_A is the marginal of μ on Ω_A. Let X_A denote the restriction of X on $(\Omega_A, \mathcal{B}_{\Omega_A})$. It is clear that $\mathcal{B}_\Omega \cup_A \mathcal{B}_{\Omega_A}$, where A ranges over the finite subsets of S. Let us denote by $\mathcal{B}_\infty = \bigcap_A \mathcal{B}_{\Omega_{S \setminus A}}$ as A ranges over the finite subsets of S. This is the algebra at infinity. We assume that we have a σ-finite positive measure η defined on $(\Sigma, \mathcal{B}_\Sigma)$. Hence we have a σ-finite positive measure $\eta^{\otimes A}$ on $(\Omega_A, \mathcal{B}_{\Omega_A})$. Now if μ is a measure on $(\Omega_A, \mathcal{B}_{\Omega_A})$ which is absolutely continuous w.r. to $\eta^{\otimes A}$, then we denote the Radon-Nikodym derivative by $\mu(x) = \frac{d\mu}{d\eta^A}(x)$.

Since we are operating in the context of countable products of Borel spaces, the existence of conditional probabilities and regular conditional probabilities is guaranteed. We denote by

$$\mu_A(B|\cdot) = E_\mu[{}^1{}_B | \mathcal{B}_{\Omega_{S \setminus A}}]$$

where ${}^1{}_B$ is the characteristic function of $B \in \mathcal{B}_{\Omega_A}$ and $E_\mu[B|\cdot]$ denotes the conditional expectation. Thus we obtain the probability kernel

$$\begin{aligned} \mu_A(\cdot|\cdot) &: \mathcal{B}_{\Omega_A} \times \Omega_{S \setminus A} \to [0, 1] \\ &: (B, x) \to \mu_A(B|x) = E_\mu[{}^1{}_B | x]. \end{aligned}$$

From the properties of conditional expectations

$$A \subseteq A' \Rightarrow \mu_{A'}(\cdot|\cdot)\mu_A(\cdot|\cdot) = \mu_{A'}(\cdot|\cdot)$$

We now come to the important definition.

DEFINITION 3.1. *A family* Π *of probability kernels* $\Pi_A(\cdot|\cdot) : \mathcal{B}_{\Omega_{A_i}} \times \Omega_{S \setminus A} \to [0, 1]$ *as* A *ranges over the finite subsets of* S *is said to be consistent of*

$$A \subseteq A' \Rightarrow \Pi_{A'}(\cdot|\cdot)\Pi_A(\cdot|\cdot) = \Pi_{A'}(\cdot|\cdot)$$

A consistent family of kernels is called a specification.

The fundamental problem posed by Dobrushin is: Given a specification $\Pi = (\Pi_A)$, A ranging over the finite subsets of S does there exist a probability measure μ on $(\Omega, \mathcal{B}_\Omega)$ such that

$$\mu_A(\cdot|\cdot) = \Pi_A(\cdot|\cdot) \; \forall \; A \text{ ranging over the finite subsets of } S. \tag{3.1}$$

Equivalently we want to solve the equations:

$$E_\mu[\cdot|\mathcal{B}_{\Omega_{S\setminus A}}] = \Pi_A, \forall \, A \subset S$$

which are finite for μ.

Equivalently, we are required to find the set of probability measures μ on $(\Omega, \mathcal{B}_\Omega)$ such that

$$\mu(B) = \int_\Omega \Pi_A(B|y)\, d\mu(y), \text{ where } B \in \mathcal{B}_\Omega, \tag{3.2}$$

A is a finite subset of S and $y \in \Omega_{S\setminus A}$.

We now introduce an equivalent formulation of Dobrushin's problem which has analogies to multi-scale methods.

Let $P(\mathcal{B}_\Omega)$ denote the set of probability measures on $(\Omega, \mathcal{B}_\Omega)$ and $\mathcal{M}(\mathcal{B}_\Omega)$ denote the set of bounded positive measures on $(\Omega, \mathcal{B}_\Omega)$. Introduce the restriction map

$$\begin{aligned} R_A \; &: \mathcal{M}(\mathcal{B}_\Omega) \to \mathcal{M}(\mathcal{B}_{\Omega_{S\setminus A}}), \text{ all } A \subset S, \text{ finite} \\ &: \mu \to R_A(\mu) \end{aligned}$$

such that for all $B \in \mathcal{B}_{\Omega_{S\setminus A}}$

$$R_A(\mu)(B) = \mu(B).$$

Introduce also the extension map associated with a specification Π by

$$\begin{aligned} T_A \; &: \mathcal{M}(\mathcal{B}_{\Omega_{A\setminus S}} \to \mathcal{M}(\mathcal{B}_\Omega) \\ &: \nu \to T_A(\nu) \end{aligned}$$

such that for all $B \in \mathcal{B}_\Omega$

$$T_A(\nu)(B) = \int \Pi_A(B|x)\, d\nu(x), \text{ all } A \subset S \text{ finite.}$$

THEOREM 3.1. *If $\mu \in \mathcal{P}(B_\Omega)$, then μ is a solution to Dobrushin's problem iff $\forall A \subset S$, finite,*

$$\mu = T_A R_A(\mu) \tag{3.3}$$

Proof. If μ is a solution to Dobrushin's problem then (3.3) is equivalent to (3.2).

Conversely, let μ satisfy (3.3). Let $B \in \mathcal{B}_{\Omega_{S\setminus A}}$ and let $B' \in \mathcal{B}_\Omega$. Note the following linearity properties of R_A and S_A.

i) $R_A(f\mu) = E_\mu(g|\mathcal{B}_{\Omega_{S\setminus A}})R_A(\mu)$, g is a positive function in $L^1(\mu)$.

ii) $T_A(f\nu) = f \circ T_A(\nu)$, where $\nu \in \mathcal{M}(\mathcal{B}_{\Omega_A})$ and $f \geq 0 \in L^1(\nu)$.

Now

$$1_B\mu = 1_B T_A R_A(\mu) = T_A(1_B R_A(\mu)).$$

Therefore,

$$\begin{aligned}\mu(B \cap B') &= 1_B\mu(B) = T_A(1_B R_A(\mu))(B') \\ &= \int \Pi_A(B'|x) d[1_B R_A(\mu)](x) \\ &= \int_B \Pi_A(B'|x) d[R_A(\mu)](x) \\ &= \int_B \Pi_A(B'|x)\, d\mu(x) \text{ since } B \in B_{\Omega_{S\setminus A}}\end{aligned}$$

■

The specifications we are interested in are called Gibbs specifications and they are given by interaction potentials. For every finite subset $A \subset S$, we choose a mapping

$$\begin{aligned}V_A &: \Omega_A \to \mathbf{R} \\ &: X_A \to V_A(X_A)\end{aligned}$$

We call V_A an interaction potential. It is said to be a pair interaction if Card $(A) > 2 \Rightarrow V_A(\cdot) = 0$. We are in particular interested in pairwise quadratic interactions which may be written as

$$V_{\{i,j\}}(X_{\{i,j\}}) = J_{ij}X(i)X(j),$$

where J_{ij}, is a real constant.

An *external field* can be introduced as:

$$V_{\{i\}}(X(i)) = h_i X(i), \text{ where } h_i \in \mathbf{R}.$$

The energy on a finite subset of pairwise interaction potentials is given by

$$H_A(X_A) = \sum_{\{i,j\}\subset A} J_{ij}X(i)X(j) + \sum_{i\in A} h_i X(i).$$

The interaction V is said to be of finite range if $\exists\ r$, such that diam $(V) > r \Rightarrow V_A(x_A) = 0$.

In general, the energy on a finite subset $A \subset S$ is given by

$$H_A(X_A) = \sum_{A'\subset A} V_{A'}(X_{A'}) \tag{3.4}$$

It turns out that this relation can be inverted which is a consequence of the Möbius Inversion formula:

Let $\mathcal{P}_f(S)$ denote the set of all finite subsets of S and let Φ and Ψ be set functions on $\mathcal{P}_f(S)$. Then

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

if and only if

$$\psi(A) = \sum_{B\subset A} \Phi(B)$$

Applying the Möbius inversion formula one can recover $V_{A'}(X_{A'})$ from $H_A(X_A)$ by the formula

$$V_A(X_A) = \sum_{A' \subset A} (-1)^{|A-A'|} H_{A'}(X_{A'}). \tag{3.5}$$

We now come to the definition of a Gibbs specification. For a $A \in \mathcal{P}_f(S)$ if $x = (y, z)$ with $y \in \Omega_A$ and $z \in \Omega_{S \setminus A}$ we denote by

$$H_A(y|z) = H_A(x)$$

Let $A \in P_f(S)$ and $Z \in \Omega_{S \setminus A}$ be such that

$$Z_A(z) = \int \exp[-H_A(y|z)dy < \infty$$

DEFINITION 3.2. *The Gibbs specification associated with a Hamiltonian H is the family of conditional probability kernels.*

$$\frac{d\pi_A}{d\eta_A}(y|z) = \frac{A}{Z_A(z)} \exp[-\frac{1}{T} H_A(y|z)].$$

It is easily checked that the above family of conditional probability kernels is consistent.

DEFINITION 3.3. *A Gibbs measure associated with the Hamiltonian H is any probability law μ on $(\Omega, \mathcal{B}_\Omega)$ such μ-almost surely*

(i) $\forall\, A \in P_f(S), Z_A(x_{S \setminus A}) < \infty$

(ii) $\frac{d\mu_A}{d\eta_A}(y|z) = \frac{1}{Z_A(z)} \exp[-\frac{1}{T} H_A(y|z)]$.

If $A \subset S$, we use the notation

$$\partial_\ell A = \{j | j \notin A, \exists i \in A \ni \rho(i,j) \leq \ell\}$$

DEFINITION 3.4. *A probability measure μ on $(\Omega, \mathcal{B}_\Omega)$ is said to ℓ-Markovian if*

$$\mu[X_A - x | X_{S \setminus A} = y] = \mu[X_A = x | X_{\partial_\ell A} = y_{\partial_\ell A}]$$

$\forall\, A \in P_f(S), \forall\, y \in \Omega_{S \setminus A}$ *and all* $x \in \Omega_A$.

DEFINITION 3.5. *μ on $(\Omega, \mathcal{B}_\Omega)$ is said to be almost Markovian if $\mu[X_A = x | X_{S \setminus A} = y]$ is continuous in y for every $A \in P_f(S)$ and every $x \in \Omega_A$.*

If μ is ℓ-Markovian then it is almost Markovian. If one is interested in almost Markovian solutions of Dobrushin's problem and if Σ is finite then it is sufficient to consider Gibbs specifications.

We end this section by citing a theorem of Dobrushin on the existence of Gibbs measures.

THEOREM 3.2. *Let Σ be finite and let Π be an almost Markovian specification. Then there exists at least one measure μ of which the Π_A's are the conditional probabilities.*

Under some technical assumptions one can show that a theorem like Theorem 3.2 holds when Σ is a Polish space.

If we denote by $\mathcal{A}$ the set of all probability measures which are solutions of Dohrushin's problem then if $\mathcal{A} \neq \phi$, then it is a convex set and under mild assumptions it can be shown to be compact and hence contains extreme points and every $\mu \in \mathcal{A}$ can be expressed

as a convex combination of these extreme points. If $\mathcal{A}$ contains only one point then the random field has no phase transitions. It is known that for any Ising model in Z^d, $d > 1$ with ferromagnetic interactions there is a critical temperature T_c above which there are no phase transitions and below which there are phase transitions.

The more general setting described here has not been used in image ϕ analysis. It would be interesting to do so.

Notes and References for Sections 2 and 3.

The exposition presented here is based on lectures given by R.L. Dobrushin at the Laboratory for Information and Decision Systems, M.I.T., in Fall 1991. For details of applications of these ideas see the M.I.T. thesis of Marroquin [3], Marroquin, Mitter, Poggio [4] and the references cited there.

References

[1] R.L. Dobrushin, Lectures at M.I.T., Fall 1991.

[2] K.R. Parthasarathy, *Probability Measures on Metric Systems*, Academic Press, New York, 1967.

[3] J.L. Marroquin, *Probabilistic Solution of Inverse Problem*, Doctoral Thesis, M.I.T., 1985.

[4] J.L. Marroquin, S. Mitter and T. Poggio, *Probabilistic Solution of Ill-posed Problems in Computer Vision*, J.A.S.A., Vol. 82, 397, 1987, pp. 76–89.

4. On Sampling Methods and Annealing Algorithms

4.1. INTRODUCTION

Discrete Markov random fields (MRF's) defined on a finite lattice have seen significant application as stochstic models for images [1], [2]. There are two fundamental problems associated with image processing based on such random field models. First, we want to generate realizations of the random fields to determine their suitability as models of our prior knowledge. Second, we want to collect statistics and perform optimizations associated with the random fields to solve model-based estimation problems, e.g., image restoration and segmentation.

According to the Hammersley-Clifford Theorem [3], (see Theorem 2.2), MRF's which are defined on a lattice are in one-to-one correspondence with Gibbs distributions. Starting with [4] there have been various constructions of Markov chains which possess a Gibbs invariant distribution, and whose common characteristic is that their transition probabilities depend only on the ratio of the Gibbs probabilities probabilities (and not on the normalization constant). These chains can be used via Monte Carlo simulation for sampling from Gibbs distributions at a fixed temperature, and for finding globally minimum energy states by slowly decreasing the temperature as in the simulated annealing (or stochastic relaxation) method [5], [6]. Certain types of diffusion processes which also have a Gibbs invariant distribution can be used for the same purposes when the random fields are continuous-valued [7], [8].

In [6], the idea of modelling an image with a compound random field for both the intensity and boundary processes was introduced. This prior random field is a MRF characterized by a Gibbs distribution. A measurement model is specified for the observed image, and the resulting posteriori random field is also a MRF characterized by a Gibbs distribution. A maximum a posteriori probability (MAP) estimate of the image based on the noisy observations is then found by minimizing the posterior Gibbs energy via simulated annealing.

Many variations and extensions of these ideas, including different estimation criteria, different methods to perform the annealing, and different methods to determine the random field parameters [9]–[12] have been used. We note that some of the alternative estimators that have been proposed do not use annealing but rather collect statistics at a fixed temperature, e.g., the maximizer of the posterior margins (MPM) and the thresholded posterior mean (TPM) estimators [9]. The scope of the MRF image models has also been enlarged over time. Most of the early work on Monte Carlo sampling methods and annealing algorithms as applied to MRF-based image processing considered finite-valued MRF's (e.g., generalized Ising models) to model discrete grey levels distributions [6]. Some more recent work has dealt with continuous- valued MRF's (e.g. Gauss-Markov models) to model continuous grey level distributions [13], [14]. In certain applications it may be advantageous to use a continuous Gauss-Markov random field model for computational and modelling considerations even when the image pixels can actually take only a finite (but large) number of grey-level values. Both Markov chain sampling methods and annealing algorithms, and diffusion-type sampling methods and annealing algorithms have been used in continuous-valued MRF-based image processing. For some of the ideas of using Gauss-Markov random fields in image processing see the paper by Moura [36] in this volume.

It should also be noted that the annealing algorithm has been used in image processing applications to minimize cost functions not derived from a MRF model (c.f. [15] for an application to edge detection), and many other non-image processing applications as well. There has been a lot of research on the convergence of discrete-state Markov chain annealing algorithms and diffusion annealing algorithms, but very few results are known about continuous-state Markov chain annealing algorithms.

Our research, described in detail in [16]–[19], addresses the following questions:

1. What is the relationship between the Markov chain sampling methods/annealing algorithms and the diffusion sampling methods/annealing algorithms?
2. What type of convergence results can be shown for discrete-time approximations of the diffusion annealing algorithms?
3. What type of convergence results can be shown for continuous-state Markov chain annealing algorithms?

In this section, we summarize some of our results. In Section [4.2] we show that continuous time interpolations of certain Markov chain sampling methods and annealing algorithms converge weakly to diffusions. In Section [4.3] we establish the convergence of a large class of discrete time modified stochastic gradient algorithms related to the diffusion annealing algorithm. Also in Section [4.4] we establish the convergence of certain continuous- state Markov chain annealing algorithms, essentially by showing that they can be expressed in the form of modified stochastic gradient algorithms. This last result gives a unifying view of the Markov chain and diffusion versions of simulated annealing algorithms. In Section [5] we briefly examine some directions for further work.

4.2. CONVERGENCE OF MARKOV CHAIN SAMPLING METHODS AND ANNEALING ALGORITHMS TO DIFFUSION

In this section we analyze the dynamics of a class of continuous state Markov chains which arise from a particular implementation of the Metropolis and the related "Heat Bath" Markov chain sampling methods [20]. Other related sampling methods (c.f. [21]) can be analyzed similarly. We show that certain continuous time interpolations of the Metropolis

and Heat Bath chains converge weakly (i.e., in distribution on path space) to Langevin diffusions. This establishes a much closer connection between the Markov chains and diffusions than just the fact that both are Markov processes which possess an invariant Gibbs distribution. We actually show that the interpolated Metropolis and Heat Bath chains converge to the same Langevin diffusion running at different time scales. This establishes a connection between the two Markov chain sampling methods which is, in general, not well understood. Our results apply to both (fixed temperature) sampling methods and (decreasing temperature) annealing algorithms.

We start by reviewing the discrete-state Metropolis and Heat Bath Markov chain sampling methods. Assume that the state space Σ is countable. Let $U(\cdot)$ be the real-valued energy function on Σ for the system. Also let T be the (positive) temperature of the system. Let $q(i,j)$ be a stationary transition probability from i to j for $i,j \in \Sigma$. The general form of the transition probability from i to j for the discrete-state Markov chains $\{X_k\}$ we consider is given by

$$p(i,j) = q(i,j)s(i,j) + m(i)1(j=i), \tag{4.1}$$

where

$$m(i) = 1 - \sum_j q(i,j)s(i,j), \tag{4.2}$$

$s(i,j)$ is a weighting factor ($0 \le s(i,j) \le 1$), and $1(\cdot)$ is an indicator function. Let $[a]_+$ denote the positive part of a, i.e., $[a]_+ = \max\{a, 0\}$. The weighting factor $s(i,j)$ is given by

$$s_M(i,j) = \exp(-[U(j) - U(i)]_+/T) \tag{4.3}$$

for the Metropolis Markov chain, and by

$$s_H(i,j) = \frac{\exp(-(U(j) - U(i))/T)}{1 + \exp(-U(j) - U(i))/T)} \tag{4.4}$$

for the Heat Bath Markov chain.

Let

$$\pi(i) = \frac{1}{Z}\exp(-U(i)/T), \quad i \in \Sigma; \quad Z = \sum_i \exp(-U(i)/T)$$

(assume $Z < \infty$). If the stochastic matrix $Q = [q(i,j)]$ is symmetric and irreducible then the detailed balance equation

$$\pi(i)p(i,j) = \pi(j)p(j,i), \quad i,j \in \Sigma,$$

is satisfied, and it follows easily that $\pi(i), i \in \Sigma$, are the unique stationary probabilities for both the Metropolis and Heat Bath Markov chains. Hence these chains may be used to sample from and to compute mean values of functionals with respect to a Gibbs distribution with energy $U(\cdot)$ and temperature T [22]. The Metropolis and Heat Bath chains can be interpreted (and simulated) in the following manner. Given the current state $X_k = i$, generate a candidate state $\tilde{X}_k = j$ with probability $q(i,j)$. Set the next state $X_{k+1} = j$ if $s(i,j) > \Theta_k$, where Θ_k is an independent random variable uniformly distributed on the interval $[0,1]$; otherwise set $X_{k+1} = i$.

We can generalize the discrete state Markov chain sampling methods described above to a continuous d- dimensional Euclidean state space as follows. Let $U(\cdot)$ be a smooth real-valued energy function on $\Sigma = \mathbf{R}^d$, and let T be the (positive) temperature. Let $q(x, y)$ be a stationary transition density from x to y for $x, y \in \mathbf{R}^d$. The general form of the transition probability density for the continuous-state Markov chain $\{X_k\}$ we consider is given by

$$p(x, y) = q(x, y)s(x, y) + m(x)\delta(y - x), \tag{4.5}$$

where

$$m(x) = 1 - \int q(x, y)s(x, y)\, dy \tag{4.6}$$

$s(i, j)$ is a weighting factor ($0 \leq s(i, j) \leq 1$), and $\delta(\cdot)$ is a Dirac-delta function. Here $s(\cdot, \cdot) = s_M(\cdot, \cdot)$ and $s(\cdot, \cdot) = s_H(\cdot, \cdot)$ (see (4.3), (4.4)) for the generalized Metropolis and Heat Bath chains, respectively.

The continuous state Metropolis and Heat Bath Markov chains can be interpreted (and simulated) analogously to the discrete state versions. In particular $q(x, y)$ is a conditional probability density for generating a candidate state $\tilde{X}_k = y$ given the current state $X_k = x$. For our analysis we shall consider the case where only a single component of the current state is changed to generate the candidate state, and the component is selected at random with all components equally likely. Furthermore, we shall require that the candidate value of the selected component depend only on the current value of the selected component. Let x_i denote the i^{th} component of the vector $x \in \mathbf{R}^d$. Let $r(x_i, y_i)$ be a transition density from x_i to y_i for $x_i, y_i \in \mathbf{R}$. Hence we set

$$q(x, y) = \frac{1}{d} \sum_{i=1}^{d} s(x, y)r(x_i, y_i) \prod_{j \neq i} \delta(y_j - x_j) \tag{4.7}$$

Suppose we take

$$r(x_i, y_i) = 1(x_i = -1)\delta(y_i - 1) + 1(x_i = 1)\delta(y_i + 1) \tag{4.8}$$

In this case, if the i^{th} coordinate of the current state X_k is selected (at random) to be changed in generating the candidate state $\tilde{X}_k$, then $\tilde{X}_{k,i}$ is ± 1 when $X_{k,l}$ is ∓ 1. If, in addition,

$$U(x) = -\sum_{j \neq i} J_{ij} x_i x_j, \quad x \in \mathbf{R}^d$$

then $\{X_k\}$ corresponds to a discrete-time kinetic Ising model with interaction energies J_{ij} [20].

Suppose instead we take

$$r(x_i, y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-(y_i - x_i)^2/2\sigma^2] \tag{4.9}$$

In this case, if the i^{th} coordinate of the current state X_k is selected (at random) to be changed in generating the candidate state $\tilde{X}_k$, the $\tilde{X}_{k,i}$ is conditionally Gaussian with mean $X_{k,i}$ and variance σ^2. In the sequel, we shall show that a family of interpolated Markov chains of this type converges (weakly) to a Langevin diffusion.

For each $\varepsilon > 0$ let $r_\varepsilon(\cdot, \cdot)$ denote the transition density in (4.9) with $\sigma^2 = \varepsilon$, and let $p_\varepsilon(\cdot, \cdot)$ denote the corresponding transition density in (4.5)–(4.7). Let $\{X_k^\varepsilon\}$ denote the

Markov chain with transition density $p_\varepsilon(\cdot,\cdot)$ and initial condition $X_0^\varepsilon = X_0$. Interpolate $\{X_k^\varepsilon\}$ into a continuous-time process $\{X^\varepsilon(t), t \le 0\}$ by setting

$$X^\varepsilon(t) = X^\varepsilon_{[t/\varepsilon]}, \quad t \le 0$$

where $[a]$ is the largest integer less than or equal to a. Now the precise definition of the weak convergence of the process $X^\varepsilon(\cdot)$ to a process $X(\cdot)$ (as $\varepsilon \to 0$) is given in [23]. The significance of the weak convergence is that it implies not only the convergence of the multivariate distribution, but also the convergence of the distributions of many interesting path functionals such as maxima, minima, and passage times (see [23] for a full discussion). To establish weak convergence here we require the following condition on $U(\cdot)$:

(A) $Y(\cdot)$ is continuously differentiable, and $\nabla U(\cdot)$ is bounded and Lipshitz continuous.

THEOREM 4.1. *Assume (A). Then there is a standard d-dimensional Wiener process $W(\cdot)$ and a process $X(\cdot)$ (with $X(0) = X_0$ in distribution, nonanticipative with respect to $W(\cdot)$, such that $X^\varepsilon(\cdot) \to X(\cdot)$ weakly as $\varepsilon \to 0$, and*

a) for the Metropolis method

$$dX(t) = \frac{\nabla U(X(t))}{2T}dt + dW(t) \tag{4.10}$$

b) for the Heat Bath method

$$dX(t) = -\frac{\nabla U(X(t))}{4T}dt + dW(t) \tag{4.11}$$

Proof. see [16] ∎

Note that Theorem 4.1 justifies our claim that the interpolated Metropolis and Heat Bath chains converge to Langevin diffusions running at different time scales. Indeed, suppose $Y(\cdot)$ is a solution of the Langevin equation

$$dY(t) = -\nabla U(Y(t))dt + \sqrt{dT}dW(t) \tag{4.12}$$

with $Y(0) = X_0$ in distribution. Then for $r(t) = t/2T$, $Y(r(\cdot))$ has then same multivariate distributions as $X(\cdot)$ satisfying (4.10), while for $r(t) = t/4T$, $Y(r(\cdot))$ has the same multivariate distributions as $X(\cdot)$ satisfying (4.11). Observe that the limit diffusion (4.10) for the Metropolis chain runs at twice the rate of the limit diffusion (4.11) for the Heat Bath chain, independent of the temperature.

To obtain Markov chain annealing algorithms we simply replace the fixed temperature T in the above Markov chain sampling methods by a temperature schedule $\{T_k\}$ (where typically $T_k \to 0$). We can establish a weak convergence result for a nonstationary continuous state Markov chain of this type as follows. Suppose $T(\cdot)$ is a positive continuous function on $[0,\infty)$. For $\varepsilon > 0$ let

$$T_k^\varepsilon = (T(k\varepsilon), \quad k = 0, 1, \ldots$$

and let $\{X_k^\varepsilon\}$ be as above but with temperature schedules $\{T_k^\varepsilon\}$. It can be shown that Theorem 4.1 is valid with T replaced by $T(t)$ in (4.10) and (4.11). Hence the Markov chain annealing algorithms converge weakly to time-scaled versions of the Markov diffusion annealing algorithm

$$dY(t) = -\nabla U(Y(t))dt + \sqrt{2T(t)}dW(t) \tag{4.13}$$

We remark that there has been a lot of work establishing convergence results for discrete state Markov chain annealing algorithms [6], [24]–[27], and also for the Markov diffusion annealing algorithm [7], [28], [29]. However, there are very few convergence results for continuous state Markov chain algorithms. We note that the weak convergence of a continuous state chain to a diffusion together with the convergence of the diffusion to the global minima of $U(\cdot)$ does not directly imply the convergence of the chain to the global minima of $U(\cdot)$; see [30] for a discussion of related issues. However, establishing weak convergence is an important first step in this regard. Indeed, a standard method for establishing the asymptotic (large-time) behavior of a large class of discrete-time recursive stochastic algorithms involves first proving weak convergence to an ODE limit. The standard method does not quite apply here because we have a discrete-time algorithm converging weakly to a nonstationary SDE limit. But calculations similar to those used to establish the weak convergence do in fact prove useful in ultimately establishing the convergence of continuous state Markov chain annealing algorithms, which is discussed in Section 4.3.2.

4.3. RECURSIVE STOCHASTIC ALGORITHMS FOR GLOBAL OPTIMIZATION IN R^D

4.3.1. Modified Stochastic Gradient Algorithms. In this section, we consider a class of algorithms for finding a global minimum of a smooth function $U(x), x \in \mathbf{R}^d$. Specifically, we analyze the convergence of a modified stochastic gradient algorithm

$$X_{k+1} = X_k - a_k(\nabla U(X_k) + \xi_k) + b_k W_k, \tag{4.14}$$

where $\{\xi_k\}$ is a sequence of $\mathbf{R}^d$-valued random variables, $\{W_k\}$ is a sequence of standard d-dimensional independent Gaussian random variables, and $\{a_k\}$, $\{b_k\}$ are sequences of positive numbers with a_k, $b_k \to 0$. An algorithm of this type arises by artificially adding the $b_k W_k$ term (via a Monte Carlo simulation) to a standard stochastic gradient algorithm

$$Z_{k+1} = Z_k - a_k(\nabla U(Z_k) + \xi_k). \tag{4.15}$$

Algorithms like (4.15) arise in a variety of optimization problems including adaptive filtering, identification and control; here the sequence $\{\xi_k\}$ is due to noisy or imprecise measurements of $\nabla U(\cdot)$ (c.f. [31]). The asymptotic behavior of $\{Z_k\}$ has been much studied. Let S and S^* be the set of local and global minima of $U(\cdot)$, respectively. It can be shown, for example, that if $U(\cdot)$ and $\{\xi_k\}$ are suitably behaved, $a_k = A/k$ for k large, and $\{Z_k\}$ is bounded, then $Z_k \to S$ as $k \to \infty$ w.p.1. However, in general $Z_k \not\to S^*$ (unless of course $S = S^*$). The idea behind adding the additional $b_k W_k$ term in (4.14) compared with (4.15) is that if b_k tends to zero slowly enough, then possibly $\{X_k\}$ (unlike $\{Z_k\}$) will avoid getting trapped in a strictly local minimum of $U(\cdot)$ (this is the usual reasoning behind simulated annealing type algorithms). We shall infact show that if $U(\cdot)$ and $\{\xi_k\}$ are suitably behaved, $a_k = A/k$ and $b_k^2 = B/k \log\log k$ for k large with $B/A > C_0$ (where C_0 is a positive constant which depends only on $U(\cdot)$), and $\{X_k\}$ is tight, then $X_k \to S^*$ as $k \to \infty$ in probability. We also give a condition for the tightness of $\{X_k\}$. We note that the convergence of Z_k to S can be established under very weak conditions on $\{\xi_k\}$ assuming $\{Z_k\}$ is bounded. Here the convergence of X_k to S^* is established under somewhat stronger conditions on $\{\xi_k\}$ assuming that $\{X_k\}$ is tight (which is weaker than boundedness).

The analysis of the convergence of $\{X_k\}$ is usually based on the asymptotic behavior of the associated ordinary differential equation (ODE)

$$\dot{z}t = -\nabla U(z(t)) \tag{4.16}$$

(c.f. [31], [32]). This motivates our analysis of the convergence of $\{X_k\}$ based on the asymptotic behavior of the associated stochastic differential equation (SDE)

$$dY(t) = -\nabla U(Y(t))dt + c(t)dW(t), \tag{4.17}$$

where $W(\cdot)$ is a standard d-dimensional Wiener process and $c(\cdot)$ is a positive function with $c(t) \to 0$ as $t \to \infty$. This is just the diffusion annealing algorithm discussed in Section 2 (see (4.13)) with $T(t) = c^2(t)/2$. The asymptotic behavior of $Y(t)$ as $t \to \infty$ has been studied intensively by a number of researchers. In [7], [29] convergence results where obtained by considering a version of (4.17) with a reflecting boundary; in [28] the reflecting boundary was removed. Our analysis of $\{X_k\}$ is based on the analysis of $Y(t)$ developed in [28] where the following result is proved: if $U(\cdot)$ is well-behaved and $c^2(t) = C/\log t$ for t large with $C > C_0$ (the same constant C_0 as above) then $Y(t) \to S^*$ as $t \to \infty$. To see intuitively how $\{X_k\}$ and $Y(\cdot)$ are related, let $t_k = \sum_{n=0}^{k-1} a_n$, $a_k = A/k$, $b_k^2 = B/k \log\log k$, $c^2(t) = C/\log t$, and $B/A = C$. Note that $b_k \sim c(t_k)\sqrt{a_k}$. Then we should have that

$$\begin{aligned} Y(t_{k+1}) &\simeq Y(t_k) - (t_{k+1} - t_k)\nabla U(Y(t_k)) + c(t_k)(W(t_{k+1}) - W(t_k)) \\ &= Y(t_k) - a_k \nabla U(Y(t_k)) + c(t_k)\sqrt{a_k V_k} \\ &\simeq Y(t_k) - a_k \nabla U(Y(t_k)) + b_k V_k \end{aligned}$$

where $\{V_k\}$ is a sequence of standard d-dimensional independent Gaussian random variables. Hence (for $\{\xi_k\}$ small enough) $\{X_k\}$ and $\{Y(t_k)\}$ should have approximately the same distributions. Of course, this is a heuristic; there are significant technical difficulties in using $Y(\cdot)$ to analyze $\{X_k\}$ because we must deal with long time intervals and slowly decreasing (unbounded) Gaussian random variables.

An algorithm like (4.14) was first proposed and analyzed in [29]. However, the analysis required that the trajectories of $\{X_k\}$ lie within a fixed ball (which as achieved by modifying (4.14) near the boundary of the ball). Hence such a version of (4.14) is only suitable for optimizing $U(\cdot)$ over a compact set. Furthermore the analysis also required ξ_k to be zero in order to obtain convergence. In our first analysis of (4.14) in [17] we also required that the trajectories of $\{X_k\}$ lie in a compact set. However, our analysis did not require ξ_k to be zero, which has important implications when $\nabla U(\cdot)$ is not measured exactly. In our later analysis of (4.14) in [18] we removed the requirement that the trajectories of $\{X_k\}$ lie in a compact set. From our point of view this is the most significant difference between our work in [18] and what is done in [29], [17] (and more generally in other work on global optimization such as [33]): we deal with unbounded processes and establish the convergence of an algorithm which finds a global minimum of a function when it is not specified a priori what bounded region contains such a point.

We now state the simplest result from [18] concerning the convergence of the modified stochastic gradient algorithm (4.14). We will require

$$a_k - \frac{A}{k}, \quad b_k = \frac{\sqrt{b}}{\sqrt{k \log\log k}}, \quad k \text{ large.} \tag{4.18}$$

and the following conditions:

(A1) $U(\cdot)$ is a C^2 function from $\mathbf{R}^d$ to $[0, \infty)$ such that the $S^* = \{x : U(x) \leq U(y)\ \forall\ y\} \neq \oslash$. (We also require some mild regularity conditions on $U(\cdot)$; see [18]).

(A2) $\underline{\lim}_{x\to\infty}\frac{|\nabla U(x)|}{|x|} > 0, \overline{\lim}_{x\to\infty}\frac{|\nabla U(x)}{|x|} < \infty.$

(A3) $\lim_{x\to\infty}\left\langle \frac{\nabla U(x)}{|\nabla U(x)|}, \frac{x}{|x|}\right\rangle = 1$

(A4) For $k = 0, 1, \ldots$, let $\mathcal{F}_k$ be the σ-field generated by $X_0, W_0, \ldots, W_{k-1}, \xi_0, \ldots, \xi_{k-1}$. There exists an $L \geq 0$, $\alpha > -1$, and $\beta > 0$ such that

$$E\{|\xi_k|^2|\mathcal{F}_k\} \leq La_k^{\alpha}(|X_k|^2 + 1), \quad |E\{\xi_k|\mathcal{F}_k\}| \leq La_k^{\beta}(|X_k| + 1) \text{ w.p. } 1$$

and W_k is independent of $\mathcal{F}_k$.

THEOREM 4.2. *Assume (A1)-(A4) hold. Let $\{X_k\}$ be given by (4.14). Then there exists a constant C_0 such that for $B/A > C_0$*

$$X_k \to S^* \text{ as } k \to \infty$$

in probability.

Proof. See [18]. ∎

Remarks:

1. The constant C_0 plays a critical role in the convergence of X_k as $k \to \infty$ and also $Y(t)$ as $t \to \infty$. In [28] it is shown that the constant C_0 (denoted there by c_0) has an interpretation in terms of the action functional for a family of perturbed dynamical systems; see [28] for a further discussion of C_0 including some examples.
2. It is possible to modify (4.14) in such a way that only the lower bound and not the upper bound on $|\nabla U(\cdot)|$ in (A2) is needed (see [18]).
3. In [18] we actually separate the problem of convergence of $\{X_k\}$ into two parts: one to establish tightness and another to establish convergence given tightness. This is analogous to separating the problem of convergence of $\{X_k\}$ into two parts: one to establish boundedness and another to establish convergence given boundedness (c.f. [31]). Now in [18] the conditions given for tightness are much stronger than the conditions given for convergence assuming tightness. For a particular algorithm it is often possible to prove tightness directly, resulting in somewhat weaker conditions than those given in Theorem 3.1.

4.3.2. Continuous-State Markov Chain Algorithm. In this section we examine the convergence of a class of continuous-state Markov chain annealing algorithms similar to those described in Section 4.2. Our approach is to write such an algorithm in the form of a modified stochastic gradient algorithm of (essentially) the type considered in Section 4.3.1. A convergence result is obtained for global optimization over all of $\mathbf{R}^d$. Some care is necessary to formulate a Markov chain with appropriate scaling. It turns out that writing the Markov chain annealing algorithm is (essentially) the form (4.14) is rather more complicated than writing standard variations of gradient algorithms which use some type of (possibly noisy) finite difference estimate of $\nabla U(\cdot)$ in the form (4.15) (c.f. [31]). Indeed, to the extend that the Markov chain annealing algorithm uses an estimate of $\nabla U(\cdot)$, it does so in a much more subtle manner than a finite difference approximation.

Although some numerical work has been performed with continuous-state Markov chain annealing algorithm [13], [14], there has been very little theoretical analysis, and furthermore the analysis of the continuous state case does not follow from the finite state case

in a straightforward way (especially for an unbounded state space). The only analysis we are aware of its in [13] where a certain asymptotic stability property is established. Since our convergence results for the continuous state Markov chain annealing algorithm are ultimately based on the asymptotic behavior of the diffusion annealing algorithm, our work demonstrates and exploits the close relationship between the Markov chain and diffusion versions of simulated annealing.

We shall perform our analysis of continuous state Markov chain annealing algorithms for a Metropolis type chain. We remark that convergence results for other continuous-state Markov chain sampling method-based annealing algorithms (such as the Heat Bath method) can be obtained by a similar procedure. Recall that the 1-step transition probability density for a continuous state Metropolis-type (fixed temperature) Markov chain is given by (see equations (4.3), (4.5), (4.6))

$$p(x,y) = q(x,y)s(x,y) + m(x)\delta(y-x)$$

where

$$m(x) = 1 - \int q(x,y)s(x,y)dy$$

and

$$s(x,y) = \exp(-[U(y) - U(x)]_+/T).$$

Here we have dropped the subscript on the weighting factor $s(x,y)$. If we replace the fixed temperature T by a temperature sequence $\{T_k\}$ we get a Metropolis-type annealing algorithm.

Our goal is to express the Metropolis-type annealing algorithm as a modified stochastic gradient algorithm like (4.14) so as to establish its convergence. This leads us to choosing a nonstationary Gaussian transition density

$$q_k(x,y) = \frac{1}{(2\pi b_k^2 \sigma^2(x))^{d/2}} \exp(\frac{|y-x|^2}{2b_k^2\sigma^2(x)}) \tag{4.19}$$

$$T_k(x) = \frac{b_k^2 \sigma_k^2(x)}{2a_k} \tag{4.20}$$

where $\sigma_k(x) = (\delta_k|x|)v^1, \delta_k \downarrow 0$.

With these choices the Metropolis-type annealing algorithm can be expressed as

$$X_{k+1} = X_k - \alpha_k(\nabla U(X_k) + \xi_k) + b_k\sigma(X_k)W_k \tag{4.21}$$

for appropriately behaved $\{\xi_k\}$. Note that (4.21) is not identical to (4.14) (because $\sigma(x) \not\equiv 1$), but is turns out that Theorem 4.2 holds for $\{X_k\}$ generated by either (4.14) or (4.21). We remark that the state dependent term $\sigma(x)$ term in (4.19) and (4.20) produces a drift toward the origin proportional to $|x|$, which is needed to establish tightness of the annealing chain.

This discussion leads us to the following continuous- state Metropolis-type annealing algorithm. Let $N(m,\Lambda)$ denote d-dimensional normal measure with mean m and covariance matrix Λ.

4.4. CONTINUOUS-STATE METROPOLIS-TYPE ANNEALING ALGORITHM:

Let $\{X_k\}$ be a Markov chain with 1 step transition probability at time k given by

$$P\{X_{k+1} \in A | X_k = x\} = \int_A s_k(x,y) dN(x, b_k^2 \sigma_k^2(x) I)(y) + m_k(x) 1_A(x) \tag{4.22}$$

where

$$m_k(x) = 1 - \int s_k(x,y) dN(x, b_k^2 \sigma^2(x) I)(y) \tag{4.23}$$

$$\sigma_k(x) = (a_k^r r|x|) V 1 \tag{4.24}$$

$$s_k(x,y) = \exp(-\frac{2a_k[U(y) - U(x)]+}{b_k^2 \quad \sigma_k^2(x)}) \tag{4.25}$$

A convergence result similar to the previous theorem can be proved for the Metropolis type annealing algorithms 19].

5. Conclusions

Monte Carlo sampling methods and annealing algorithms have found significant application to MRF-based image processing. These algorithms fall broadly into two groups: Markov chain and diffusion methods. The discrete-state Markov chain algorithms have been used with finite range MRF models, while both continuous-state Markov chain and diffusion algorithms have been used with continuous range MRF models. We note that there are some very interesting questions related to the parallel implementation of these Monte Carlo procedures which we have not discussed here: see [34].

It seems to us that some experimental comparisons of continuous state Markov chain and diffusion-type annealing algorithms (practically implemented by the modified stochastic gradient algorithms described above) on image segmentation and restoration problems would be of some interest. We are not aware of any explicit comparisons of this type in the literature. It might also be useful to examine the application of the modified stochastic gradient algorithms to adaptive pattern recognition, filtering and identification, where stochastic gradient algorithms are frequently employed. Because of the slow convergence of the modified stochastic gradient algorithms, offline applications will probably be required. One particular application which might prove fruitful is training multilayer feedforward "neural nets", which is a nonconvex optimization problem often plagued with local minima. A rigorous analysis which discusses the learning problem for Boltzmann machines has been carried out in [35] by viewing it as a Maximum Likelihood estimation. In this paper however convergence to a global maximum is not proved and it would be interesting to see whether the ideas of [19] can be used to do this.

Acknowledgements

This research has been supported by the Air Force Office of Scientific Research under grant AFOSR-89-0276-C and the Army Research Office under grant DAAL03-92-G-0115.

References

[1] R.L. Kashyap, R. Chellappa, *Estimation and Choice of Neighbors in Spatial Interaction Models of Images*, IEEE Trans. on Info. Theory, Vol. 29, 1983, p. 60–72.

[2] J.W. Woods, *Two-Dimensional Discrete Markovian Fields*, IEEE Trans. Inf. Theory, Vol. 18, 1972, p. 232–240.

[3] J. Besag, *Spatial Interaction and the Statistical Analysis of Lattice Systems*,, J. Royal Stat. Soc., Vol. 34, 1972, p. 75–83.

[4] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, *Equation of State Calculations by Fast Computing Machines*, J. Phys. Chem., Vol. 21, No. 6, 1953, p. 1087.

[5] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi, *Optimization by Simulated Annealing*, Science, Vol. 220, 1983, p. 671–680.

[6] S. Geman and D. Geman, *Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images*, IEEE Trans. Pattern Anal. and Machine Intell., Vol. 6, 1984, p. 721–741.

[7] S. Geman and C.R. Hwang, *Diffusions for Global Optimization*, SIAM Journal Control and Optimization, Vol. 24, 1986, p. 1031–1043.

[8] U. Grenander, *Tutorial in Pattern Theory*, Div. of Applied Math, Brown University, 1984.

[9] J.L. Marroquin, S. Mitter, T. Poggio, *Probabilistic Solution of Ill-Posed Problems in Computational Vision*, J. Amer. Statist. Assoc., Vol. 82, 1987, p. 76-89.

[10] B. Gidas, *A Renormalization Group Approach to Image Processing Problems*, IEEE Trans. on Pattern Anal. and Machine Intell., Vol. PAMI-11, February 1989, p. 164–180.

[11] S. Lakshmanan, and H. Derin, *Simultaneous Parameter Estimation and Segmentation of Gibbs Random Fields Using Simulated Annealing*, IEEE Trans. on Pattern Anal. and Machine Intell., Vol. PAMI-11, No. 8, August 1989, p. 799–813.

[12] D. Geman, S. Geman, C. Graffigne, and P. Dong, *Boundary Detection by Constrained Optimization*, IEEE Trans. on Pattern Anal. and Machine Intell., Vol. PAMI-12, No. 7, July 1990, p. 609–628.

[13] F.J. Jeng and J.W. Woods, *Simulated Annealing in Compound Gaussian Random Fields*, IEEE Trans. Info. Theory, Vol. 36, No. 1, 1990, p. 94–107.

[14] T. Simchony, R. Chellappa and Z. Lichtenstein, *Relaxation Algorithms for MAP Estimation of Grey-Level Images with Multiplicative Noise*, IEEE Trans. Info. Theory, Vol. 36, No. 3, 1990, p. 608–613.

[15] H.L. Tan, S.B. Gelfand and E.J. Delp, *A Cost Minimization Approach to Edge Detection Using Simulated Annealing*, Proceedings of the IEEE Computer Vision and Pattern Recognization Conference, San Diego, CA. p. 86–91; submitted to IEEE Trans. Pattern Anal. and Machine Intell.

[16] S.B. Gelfand, S.K. Mitter, *Weak Convergence of Markov Chain Sampling Methods and Annealing Algorithms to Diffusions*, J. of Optimization Theory and Applications, Vol. 68, No. 3, March 1991.

[17] S.B. Gelfand, S.K. Mitter, *Simulated Annealing-Type Algorithms for Multivariate Optimization*, Algorithmica, 1991, pp.419–436.

[18] S.B. Gelfand and S.K. Mitter, *Recursive Stochastic Algorithms for Global Optimization in* $\mathbf{R}^d$, SIAM Journal Control and Optimization, Vol. 29, No. 5, pp. 999–1018, September 1991.

[19] S.B. Gelfand and S.K. Mitter, *Metropolis-Type Annealing Algorithms for Global Optimization in* $\mathbf{R}^d$, SIAM Journal Control and Optimization, Nov. 1992.

[20] K. Binder, *Monte Carlo Methods in Statistical Physics*, Springer-Verlag, Berlin, 1978.

[21] W.K. Hastings, *Monte Carlo Sampling Methods Using Markov Chains and Their Applications*, Biometrika, Vol. 57, 1970, p. 97–109.

[22] K.L. Chung, *Markov Processes with Stationary Transition Probabilities*, Springer-Verlag, Heidelberg, Germany, 1960.

[23] P. Billingsley, *Convergence of Probability Measures*, John Wiley & Sons, New York, NY, 1968.

[24] B. Gidas, *Nonstationary Markov Chains and Convergence of the Annealing Algorithm*, J. of Statistical Physics, Vol. 39, 1985, p. 73–131.

[25] B. Hajek, *Cooling Schedules for Optimal Annealing*, Mathematics of Operations Research, Vol. 13, 1988, p. 311–329.

[26] D. Mitra, F. Romeo, and A. Sangiovanni-Vincentelli, *Convergence and Finite-Time Behavior of Simulated Annealing*, Advances in Applied Probability, Vol. 18, 1986, p. 747–771.

[27] J. Tsitsiklis, *Markov Chains with Rare Transitions and Simulated Annealing*, Mathematics of Operations Research, Vol. 14, 1989, p. 70–90.

[28] T.S. Chiang, C.R. Hwang, and S.J. Sheu, *Diffusion for Global Optimization in* $\mathbf{R}^n$, SIAM Journal Control and Optimization, Vol. 25, 1987, p. 737-752.

[29] H.J. Kushner, *Asymptotic Global Behavior for Stochastic Approximation and Diffusions with Slowly Decreasing Noise Effects: Global Minimization Via Monte Carlo*, SIAM Journal Applied Mathematics, Vol. 47, 1987, p. 169–185.

[30] H.J. Kushner, *Approximation and Weak Convergence Methods for Random Processes*, MIT Press, Cambridge, MA, 1984.

[31] H.J. Kushner and D. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, Berlin, Germany, 1987.

[32] L. Ljung, *Analysis of Recursive Stochastic Algorithms*, IEEE Trans. on Automatic Control, Vol. AC-22, 1977, p. 551–575.

[33] L.C. W. Dixon and G.P. Szego, *Towards Global Optimization*, North Holland, 1978.

[34] J.L. Marroquin, *Probabilistic Solution of Inverse Problems*, Ph.D. Thesis, LIDS-TH-1500, Laboratory for Information and Decision Systems, MIT, Cambridge, MA, 1985.

[35] H.J. Sussman, *On the Convergence of Learning Algorithms for Boltzmann Machines*, Tech. Rept. SYCON-88-03, Rutgers Center for Dynamical Systems, Rutgers University.

[36] J.F. deMoura: Paper in this volume.