ON SAMPLING METHODS AND ANNEALING ALGORITHMS [1]

Saul B. Gelfand
School of Electrical Engineering
Purdue University
West Lafayette, IN 47907

and

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

[1]The research of the first author has been supported by the National Science Foundation under contract ECS-8910073. The research of the second author has been supported by the Air Force Office of Scientific Research under contract 89-0276B and by the Army Research Office under contract DAAL03-86-K-0171 (Center for Intelligent Control Systems).

1 Introduction

Discrete Markov random fields (MRF's) defined on a finite lattice have seen significant application as stochastic 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], 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 (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].

Many of the fundamental ideas on MRF-based image processing stem from [6], which introduced the idea of modelling an image with a compound random field for both the intensity and boundary processes. 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.

There have been numerous variations and extensions of the ideas in [6], including different estimation criteria, different methods to perform the annealing, and different methods to determine the random field parameters [9]–[12]. 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 marginals (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.

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 we well. There has been a lot of research on the convergence of discrete-state Markov chain annealing algorithms and diffu-

sion 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 types of convergence results can be shown for discrete-time approximations of the diffusion annealing algorithms?
3. What types of convergence results can be shown for continuous-state Markov chain annealing algorithms?

In this paper, we summarize some of our results. In Section 2 we show that continuous time interpolations of certain Markov chain sampling methods and annealing algorithms converge weakly to diffusions. In Section 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 3 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 4 we briefly examine some directions for further work.

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

$$p(i,j) = q(i,j)s(i,j) + m(i)1(j=i), \tag{2.1}$$

where

$$m(i) = 1 - \sum_j q(i,j)s(i,j), \tag{2.2}$$

$s(i,j)$ is a weighting factor ($0 \leq s(i,j) \leq 1$), 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

$$s_M(i,j) = \exp(-[U(j) - U(i)]_+/T) \tag{2.3}$$

for the Metropolis Markov chain, and by

$$s_H(i,j) = \frac{\exp(-(U(j) - U(i))/T)}{1 + \exp(-(U(j) - U(i))/T)} \tag{2.4}$$

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 $q(\boldsymbol{x}, \boldsymbol{y})$ be a stationary transition density from $\boldsymbol{x}$ to $\boldsymbol{y}$ for $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{R}^d$. The general form of the transition probability density for the continuous-state Markov chain $\{X_k\}$ we consider is given by

$$p(\boldsymbol{x}, \boldsymbol{y}) = q(\boldsymbol{x}, \boldsymbol{y})s(\boldsymbol{x}, \boldsymbol{y}) + m(\boldsymbol{x})\delta(\boldsymbol{y} - \boldsymbol{x}), \tag{2.5}$$

where

$$m(\boldsymbol{x}) = 1 - \int q(\boldsymbol{x}, \boldsymbol{y})s(\boldsymbol{x}, \boldsymbol{y})d\boldsymbol{y}, \tag{2.6}$$

$s(i,j)$ 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 (2.3), (2.4)) 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 $q(\boldsymbol{x}, \boldsymbol{y})$ is a conditional probability density for generating a candidate state $\tilde{X}_k = \boldsymbol{y}$ given the current state $X_k = \boldsymbol{x}$. 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

$$q(x,y) = \frac{1}{d}\sum_{i=1}^{d} s(x,y) r(x_i, y_i) \prod_{j \neq i} \delta(y_j - x_j) \tag{2.7}$$

Suppose we take

$$r(x_i, y_i) = 1(x_i = -1)\delta(y_i - 1) + 1(x_i = 1)\delta(y_i + 1) \tag{2.8}$$

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 $X_{k,1}$ 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

$$r(x_i, y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-(y_i - x_i)^2 / 2\sigma^2\right] \tag{2.9}$$

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 (2.9) with $\sigma^2 = \varepsilon$, and let $p_\varepsilon(\cdot,\cdot)$ denote the corresponding transition density in (2.5)-(2.7). 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 $\{X^\varepsilon(t), t \leq 0\}$ by setting

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

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 2.1: 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

$$dX(t) = \frac{\nabla U(X(t))}{2T} dt + dW(t) \tag{2.10}$$

b) for the Heat Bath method

$$dX(t) = -\frac{\nabla U(X(t))}{4T}dt + dW(t) \tag{2.11}$$

Proof: see [16]

Note that Theorem 2.1 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

$$dY(t) = -\nabla U(Y(t))dt + \sqrt{2T}dW(t) \tag{2.12}$$

with $Y(0) = X_0$ in distribution. Then for $\tau(t) = t/2T$, $Y(\tau(\cdot))$ has then same multivariate distributions as $X(\cdot)$ satisfying (2.10), while for $\tau(t) = t/4T$, $Y(\tau(\cdot))$ has the same multivariate distributions as $X(\cdot)$ satisfying (2.11). Observe that the limit diffusion (2.10) for the Metropolis chain runs at twice the rate of the limit diffusion (2.11) 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 2.1 is valid with T replaced by $T(t)$ in (2.10) and (2.11). Hence the Markov chain annealing algorithms converge weakly to time-scaled versions of the Markov diffusion annealing algorithm

$$dY(t) = -\nabla U(Y(t))dt + \sqrt{2T(t)}dW(t) \tag{2.13}$$

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 3.2.

3 Recursive Stochastic Algorithms for Global Optimization in $\mathbf{R}^d$

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 modied stochastic gradient algorithm

$$X_{k+1} = X_k - a_k(\nabla U(X_k) + \xi_k) + b_k W_k, \tag{3.1}$$

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 artifically adding the $b_k W_k$ term (via aMonte Carlo simulation) to a standard stochastic gradient algorithm

$$Z_{k+1} = Z_k - a_k(\nabla U(Z_k) + \xi_k). \tag{3.2}$$

Algorithms like (3.2) 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 minimaa 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 (3.1) compared with (3.2) 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 $b_k^2 = B/kloglogk$ 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)

$$\dot{z}(t) = -\nabla U(z(t)) \tag{3.3}$$

(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)

$$dY(t) = -\nabla U(Y(t))dt + c(t)dW(t), \tag{3.4}$$

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 (2.13)) 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 (3.4) 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 $c^2(t) = C/logt$ 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 $t_k = \sum_{n=0}^{k-1} a_n$, $a_k = A/k$, $b_k^2 = B/kloglogk$, $c^2(t) = C/logt$, 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 (3.1) was first proposed and analyzed in [29]. However, the analysis required that the trajectories of $\{X_k\}$ lie within a fixed ball (which was achieved by modifying (3.1) near the boundary of the ball). Hence such a version of (3.1) 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 (3.1) 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 (3.1) 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 (3.1). We will require

$$a_k = \frac{A}{k}, \quad b_k = \frac{\sqrt{B}}{\sqrt{kloglogk}}, \quad k \text{ large.} \tag{3.5}$$

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

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

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

Theorem 3.1: Assume (A1)-(A4) hold. Let $\{X_k\}$ be given by (3.1). Then there exists a constant C_0 such that for $B/A > C_0$

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

in probability.

Proof: See [18]

Remarks:

1. The constant C_0 plays a critical role in the convergence of X_k as $k \rightarrow \infty$ and also $Y(t)$ as $t \rightarrow \infty$. 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 (3.1) 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 $\{Z_k\}$ 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.

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 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 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 in (essentially) the form (3.1) 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 (3.2) (c.f. [31]). Indeed, to the extent 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 algorithms [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 is 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 (2.3), (2.5), (2.6))

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

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 (3.1) so as to establish its convergence. This leads us to choosing a nonstationary Gaussian transition density

$$q_k(x,y) = \frac{1}{(2\pi b_k^2 \sigma^2(x))^{d/2}} \exp\left(\frac{|y-x|^2}{2b_k^2\sigma^2(x)}\right) \tag{3.6}$$

and a state-dependent temperature sequence

$$T_k(x) = \frac{b_k^2\sigma^2(x)}{2a_k} \tag{3.7}$$

where $\sigma(\cdot)$ is a continuous function from $\mathbf{R}^d$ to $[1,\infty)$ such that

$$\sigma(x) \sim |x| \text{ as } x \to \infty$$

(e.g. $\sigma(x) = 1 + |x|$ or $\sigma(x) = \max\{1,|x|\}$ will do). With these choices the Metropolis-type annealing algorithm can be expressed as

$$X_{k+1} = X_k - a_k(\nabla U(X_k) + \xi_k) + b_k\sigma(X_k)W_k \tag{3.8}$$

for appropriately behaved $\{\xi_k\}$. Note that (3.8) is not identical to (3.1) (because $\sigma(x) \not\equiv 1$), but it turns out that Theorem 3.1 holds for $\{X_k\}$ generated by either (3.1) or (3.8). We remark that the state dependent term $\sigma(x)$ term in (3.6) and (3.7) 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.

3.3 Continuous-State Metropolis-Type Annealing Algorithm:

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

$$P\{X_{k+1} \in A | X_k = x\} = \int_A s_k(x,y) dN(x, b_k^2 \sigma^2(x) I)(y) + m_k(x) 1_A(x) \tag{3.9}$$

where

$$m_k(x) = 1 - \int s_k(x,y) dN(x, b_k^2 \sigma^2(x) I)(y) \tag{3.10}$$

and

$$s_k(x,y) = \exp\left(- \frac{2a_k [U(y) - U(x)]_+}{b_k^2 \quad \sigma^2(x)} \right) \tag{3.11}$$

We now state a convergence result from [19] concerning the convergence of the continuous-state Metropolis type annealing algorithm. Let the sequences $\{a_k\}$ and $\{b_k\}$ be given by (3.5).

Theorem 3.2: Assume (A1)–(A3) hold. Let $\{X_k\}$ be the Markov chain with transition probability given by (3.9)–(3.11). Then there exists a constant C_0 such that for $B/A > C_0$

$$X_k \rightarrow S* \text{ as } k \rightarrow \infty$$

in probability.

Proof: We mentioned above that Theorem 3.1 holds for either (3.1) or (3.8). Hence it is enough to show that $\{\xi_k\}$ defined by (3.8) satisfies (A4). In [19] it is shown that there exists an $L \geq 0$ such that

$$E\{|\xi_k|^2 | \mathcal{F}_k\} \leq L \frac{b_k}{a_k} (|X_k|^2 + 1), \quad |E\{\xi_k | \mathcal{F}_k\}| \leq L \frac{a_k}{b_k} (|X_k| + 1) \quad w.p.1$$

It follows that (A4) is satisfied with $\alpha = -1/2$ and $\beta \in (0, 1/2)$. □

Remarks:

1. The constant C_0 is the same constant described in Remark 1 following Theorem 3.1
2. It is possible to modify (3.9)–(3.11) in such a way that only the lower bound and not the upper bound on $|\nabla U(\cdot)|$ from (A2) is needed (see [19]).

4 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].

In this paper we summarized some of our research which has investigated the relationship between the various Markov chain and diffusion sampling methods and annealing algorithms. We demonstrated the weak convergence of certain interpolated Markov chain sampling methods and annealing algorithms to diffusions. We also established the large-time convergence of a class of discrete-time modified stochastic gradient algorithms based on the asymptotic behavior of the associated diffusion annealing algorithm. We further established the large-time convergence of a continuous-state Markov chain annealing algorithm by writing it in the form of such a modified stochastic gradient algorithm. The convergence here is to the global minima of an energy cost function defined on the entire d-dimensional Euclidean space.

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 [35].

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, *Simulatneous 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,* to appear in J. Optimization Theory and Applications.

[17] S. B. Gelfand, S. K. Mitter, *Simulated Annealing-Type Algorithms for Multivariate Optimization,* to appear in Algorithmica.

[18] S. B. Gelfand and S. K. Mitter, *Recursive Stochastic Algorithms for Global Optimization in* $\mathbf{R}^d$, to appear in SIAM Journal Control and Optimization.

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

[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, 1978.

[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] T. Khanna, *Foundations for Neural Networks,* Addison-Wesley, Reading, MA, 1990.