SIAM J. CONTROL AND OPTIMIZATION
Vol. 31, No. 1, pp. 111-131, January 1993

METROPOLIS-TYPE ANNEALING ALGORITHMS FOR GLOBAL OPTIMIZATION IN $\mathbb{R}^d$*

SAUL B. GELFAND† AND SANJOY K. MITTER‡

Abstract. The convergence of a class of Metropolis-type Markov-chain annealing algorithms for global optimization of a smooth function $U(\cdot)$ on $\mathbb{R}^d$ is established. No prior information is assumed as to what bounded region contains a global minimum. The analysis contained herein is based on writing the Metropolis-type algorithm in the form of a recursive stochastic algorithm $X_{k+1}=X_k-a_k(\nabla U(X_k)+\xi_k)+b_kW_k$, where $\{W_k\}$ is a standard white Gaussian sequence, $\{\xi_k\}$ are random variables, and $a_k=A/k$, $b_k=\sqrt{B}/\sqrt{k\log\log k}$ for k large. Convergence results for $\{X_k\}$ are then applied from our previous work [*SIAM Journal on Control and Optimization*, 29 (1991), pp. 999-1018]. Since the analysis of $\{X_k\}$ is based on the asymptotic behavior of the related Langevin-type Markov diffusion annealing algorithm $dY(t)=-\nabla U(Y(t))\,dt+c(t)\,dW(t)$, where $W(\cdot)$ is a standard Wiener process and $c(t)=\sqrt{C}/\sqrt{\log t}$ for t large, this work demonstrates and exploits the close relationship between the Markov chain and diffusion versions of simulated annealing.

Key words. global optimization, random optimization, simulated annealing, stochastic gradient algorithms, Markov chains

AMS(MOS) subject classifications. 65K10, 90C30, 60J60

1. Introduction. Let $U(\cdot)$ be a real-valued function on some set Σ. The global optimization problem is to find an element of the set $S^*=\{x\colon U(x)\leqq U(y)$ for all $y\in\Sigma\}$ (assuming that $S^*\neq\emptyset$). Recently, there has been much interest in the simulated annealing method for global optimization. Annealing algorithms were initially proposed for finite optimization (Σ finite), and later developed for continuous optimization ($\Sigma=\mathbb{R}^d$). An annealing algorithm for finite optimization was first suggested in [17], [2] and is based on simulating a finite-state Metropolis-type Markov chain. The Metropolis algorithm and other related algorithms such as the "heat bath" algorithm, were originally developed as Markov chain sampling methods for sampling from a Gibbs distribution [1]. The asymptotic behavior of finite state Metropolis-type annealing algorithms has been extensively analyzed [3], [5], [9], [12], [14], [21], [24], [25].

A continuous-time annealing algorithm for continuous optimization was first suggested in [10], [13], and is based on simulating a Langevin-type Markov diffusion as follows:

$$dY(t)=-\nabla U(Y(t))\,dt+c(t)\,dW(t). \tag{1.1}$$

Here $U(\cdot)$ is a smooth function on $\mathbb{R}^d$, $W(\cdot)$ is a standard d-dimensional Wiener process, and $c(\cdot)$ is a positive function with $c(t)\to 0$ as $t\to\infty$. In the terminology of simulated annealing algorithms, $U(x)$ is called the energy of state x, and $T(t)=c^2(t)/2$ is called the temperature at time t. Note that for a fixed temperature $T(t)=T$, the resulting Langevin diffusion, like the Metropolis chain, has a Gibbs distribution $\propto\exp(-U(x)/T)$ as its invariant measure. Now (1.1) can be viewed as adding decreasing white Gaussian noise to the continuous time gradient algorithm

$$\dot{z}(t)=-\nabla U(z(t)). \tag{1.2}$$

* Received by the editors July 2, 1990; accepted for publication (in revised form) September 5, 1991. This research was supported by National Science Foundation contract ECS-8910073, Air Force Office of Scientific Research contract 89-0276B, and by Army Research Office contract DAAL03-86-K-0171 (Center for Intelligent Control Systems).

† School of Electrical Engineering, Purdue University, West Lafayette, Indiana 47907.

‡ Department of Electrical Engineering and Computer Science, and Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

We use (1.1) instead of (1.2) for minimizing $U(\cdot)$ to avoid getting trapped in strictly local minima. The asymptotic behavior of $Y(t)$ as $t\to\infty$ has been studied in [4], [10], [11], [18]. In [10], [18] convergence results were obtained for a version of (1.1), which was modified to constrain the trajectories to lie in a fixed bounded set (and hence is only applicable to global optimization over a compact subset of $\mathbb{R}^d$); in [4], [11] results were obtained for global optimization over all of $\mathbb{R}^d$. Chiang, Hwang, and Sheu's main result from [4] can be roughly stated as follows: If $U(\cdot)$ is suitably behaved and $c^2(t)=C/\log t$ for t large with $C>C_0$ (a constant depending only on $U(\cdot)$), then $Y(t)\to S^*$ as $t\to\infty$ in probability.

A discrete-time annealing algorithm for continuous optimization was suggested in [8], [18] and is based on simulating a recursive stochastic algorithm

$$X_{k+1}=X_k-a_k(\nabla U(X_k)+\xi_k)+b_kW_k. \tag{1.3}$$

Here $U(\cdot)$ is again a smooth function on $\mathbb{R}^d$, $\{\xi_k\}$ is a sequence of $\mathbb{R}^d$-valued random variables, $\{W_k\}$ is a sequence of independent standard d-dimensional Gaussian random variables, and $\{a_k\}$, $\{b_k\}$ are sequences of positive numbers with $a_k, b_k\to 0$ as $k\to\infty$. Algorithm (1.3) could arise from a discretization or numerical integration of the diffusion (1.1) so as to be suitable for implementation on a digital computer; in this case, ξ_k is due to the discretization error. Alternatively, algorithm (1.3) could arise by artificially adding decreasing white Gaussian noise (i.e., the b_kW_k terms) to a stochastic gradient algorithm

$$Z_{k+1}=Z_k-a_k(\nabla U(Z_k)+\xi_k), \tag{1.4}$$

which arises in a variety of optimization problems including adaptive filtering, identification and control; in this case, ξ_k is due to noisy or imprecise measurements of $\nabla U(\cdot)$ (cf. [19]). We again use (1.3) instead of (1.4) for minimizing $U(\cdot)$ to avoid getting trapped in strictly local minima. In the following, we refer to (1.4) and (1.3) as standard and modified stochastic gradient algorithms, respectively. The asymptotic behavior of X_k as $k\to\infty$ has been studied in [8], [18]. In [18] convergence results were obtained for a version of (1.3), which was modified to constrain the trajectories to lie in a compact set (and hence is only applicable to global optimization over a compact subset of $\mathbb{R}^d$); in [8] results were obtained for global optimization over all of $\mathbb{R}^d$. Also, in [18] convergence is obtained essentially only for the case where $\xi_k=0$; in [8] convergence is obtained for $\{\xi_k\}$ with unbounded variance. This latter fact has important implications when $\nabla U(\cdot)$ is not measured exactly. Our main result from [8] can be roughly stated as follows: 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 $B/A>C_0$ (the same C_0 as above), and $\{X_k\}$ is tight, then $X_k\to S^*$ as $k\to\infty$ in probability (conditions are also given in [8] for tightness of $\{X_k\}$). Our analysis in [8] of the asymptotic behavior of X_k as $k\to\infty$ is based on the behavior of the associated stochastic differential equation (SDE) (1.1). This is analogous to the well-known method of analyzing the asymptotic behavior of Z_k as $k\to\infty$ based on the behavior of the associated ordinary differential equation (ODE) (1.2) [19], [20].

It has also been suggested that continuous optimization might be performed by simulating a continuous-state Metropolis-type Markov chain [10]. This method has been applied to the restoration of noise corrupted images [16], [23]. In these works, Gaussian random field models are used so that the state space is unbounded. Although some numerical work has been performed with continuous-state Metropolis-type annealing algorithms, 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 of which we know is in [16], where a certain asymptotic stability property is established for a related algorithm and a particular cost function that arises in a problem of image restoration.

In this paper, we demonstrate the convergence of a class of continuous-state Metropolis-type Markov-chain annealing algorithms for general cost functions. Our approach is to write such an algorithm in the form of a modified stochastic gradient algorithm (1.3) for suitable choice of ξ_k, and to apply results from [8]. A convergence result is obtained for global optimization over all of $\mathbb{R}^d$. Some care is necessary to formulate a Metropolis-type Markov chain with appropriate scaling. It turns out that writing the Metropolis-type annealing algorithm in the form (1.3) is more complicated than writing standard variations of gradient algorithms, which use some type of finite-difference estimate of $\nabla U(\cdot)$, in the form (1.4) (cf. [19]). Indeed, to the extent that the Metropolis-type annealing algorithm uses an estimate of $\nabla U(\cdot)$, it does so in a much more subtle manner than a finite-difference approximation, as is seen in the analysis.

Since our convergence results for the Metropolis-type Markov-chain annealing algorithm are ultimately based on the asymptotic behavior of the Langevin-type Markov diffusion annealing algorithm, this paper demonstrates and exploits the close relationship between the Markov chain and diffusion versions of simulated annealing, which is particularly interesting in view of the fact that the development and analysis of these methods has proceeded more or less independently. We note that similar convergence results for other annealing algorithms based on the continuous-state Markov-chain sampling method (such as the "heat bath" method) can be obtained by a procedure similar to that used in this paper.

It is important to note that, although we establish the convergence of the Metropolis-type Markov-chain annealing algorithm by effectively comparing it with the Langevin-type Markov diffusion annealing algorithm, the finite-time behavior of the algorithms may be quite different. Some indication of this arises in the analysis; see Remarks 1 and 2 in § 4.

The paper is organized as follows. In § 2 we discuss appropriately modified versions of tightness and convergence results for modified stochastic gradient algorithms, as given in [8]. In § 3 we present a class of continuous-state Metropolis-type annealing algorithms and state some convergence theorems. In § 4 we prove the convergence theorems of § 3, using the results of § 2.

2. Modified stochastic gradient algorithms. In this section, we give convergence and tightness results for modified stochastic gradient algorithms of the type described in § 1. The algorithms and theorems discussed below are a slight variation on the results of [8] and are appropriate for proving convergence and tightness for a class of continuous state Metropolis-type annealing algorithms (see §§ 3 and 4).

We use the following notation throughout the paper. Let $\nabla U(\cdot)$, $\Delta U(\cdot)$, and $HU(\cdot)$ denote the gradient, Laplacian, and Hessian matrix of $U(\cdot)$, respectively. Let $|\cdot|$, $\langle\cdot,\cdot\rangle$, and $\otimes$ denote Euclidean norm, inner product, and outer product, respectively. For real numbers a and b, let $a \vee b = \text{maximum}\{a, b\}$, $a \wedge b = \text{minimum}\{a, b\}$, $[a]^+ = a \vee 0$, and $[a]^- = a \wedge 0$. For a process $\{X_k\}$ and a function $f(\cdot)$, let $E_{n,x}\{f(X_k)\}$ denote conditional expectation, given $X_n = x$, and let $E_{n_1,x_1;n_2,x_2}\{f(X_k)\}$ denote conditional expectation, given $X_{n_1} = x_1$ and $X_{n_2} = x_2$ (more precisely, these are suitably fixed versions of the conditional expectation). Also, for a measure $\mu(\cdot)$ and a function $f(\cdot)$, let

$\mu(f)=\int f\,d\mu$. Finally, let $N(m, R)(\cdot)$ denote normal measure with mean m and covariance matrix R, and let I denote the identity matrix.

2.1. Convergence. In this section, we consider the convergence of the discrete-time algorithm

$$X_{k+1}=X_k-a_k(\nabla U(X_k)+\xi_k)+b_kW_k. \tag{2.1}$$

Here $U(\cdot)$ is a smooth real-valued function on $\mathbb{R}^d$, $\{\xi_k\}$ is a sequence of $\mathbb{R}^d$-valued random variables, $\{W_k\}$ is a sequence of independent standard d-dimensional Gaussian random variables, and

$$a_k=\frac{A}{k}, \quad b_k=\frac{\sqrt{B}}{\sqrt{k\log\log k}}, \qquad k \text{ large},$$

where A, B are positive constants.

For $k=0, 1, \cdots$, let $\mathcal{F}_k=\sigma(X_0, W_0, \cdots, W_{k-1}, \xi_0, \cdots, \xi_{k-1})$. In the following, we consider the following conditions (α, β are constants whose values are specified later).

Condition 1. $U(\cdot)$ is a C^2 function from $\mathbb{R}^d$ to $[0, \infty)$ such that

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

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

$$\inf_x\,(|\nabla U(x)|^2-\Delta U(x))>-\infty.$$

Condition 2. For $\varepsilon>0$, let

$$d\pi^\varepsilon(x)=\frac{1}{Z^\varepsilon}\exp\left(-\frac{2U(x)}{\varepsilon^2}\right)dx, \qquad Z^\varepsilon=\int\exp\left(-\frac{2U(x)}{\varepsilon^2}\right)dx<\infty.$$

π^ε has a weak limit π as $\varepsilon\to 0$.

Condition 3. Let K be a compact subset of $\mathbb{R}^d$. Then there exists $L, k_0 \geqq 0$ such that, for every $k \geqq k_0$,

$$E\{|\xi_k|^2|\mathcal{F}_k\}\leqq La_k^\alpha, \qquad \forall X_k\in K, \quad \text{with probability one (w.p.1)}, \tag{2.2a}$$

$$|E\{\xi_k\,|\,\mathcal{F}_k\}|\leqq La_k^\beta, \qquad \forall X_k\in K, \quad \text{w.p.1.} \tag{2.2b}$$

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

We note that π concentrates on S^*, the global minima of $U(\cdot)$. The existence of π and a simple characterization in terms of $HU(\cdot)$ is discussed in [15].

In [4] and [8], it was shown that there exists a constant C_0, which plays a critical role in the convergence of (1.1) and (1.3), respectively (in [4] C_0 was denoted by c_0). C_0 has an interpretation in terms of the action functional for the perturbed dynamical systems

$$dY^\varepsilon(t)=-\nabla U(Y^\varepsilon(t))\,dt+\varepsilon\,dW(t). \tag{2.3}$$

Now, for $\phi(\cdot)$ an absolutely continuous function on $\mathbb{R}^d$, the (normalized) action functional for (2.3) is given by

$$I(t, x, y)=\inf_{\substack{\phi(0)=x\\ \phi(t)=y}}\frac{1}{2}\int_0^t|\dot\phi(s)+\nabla U(\phi(s))|^2\,ds.$$

According to [4],

$$C_0 = \tfrac{3}{2} \sup_{x,y \in S_0} (V(x,y) - 2U(y)),$$

where $V(x,y) = \lim_{t\to\infty} I(t,x,y)$, and S_0 is the set of all the stationary points of $U(\cdot)$, i.e., $S_0 = \{x : \nabla U(x) = 0\}$; see [4] for a further discussion of C_0, including some examples.

Let $K_1 \subset \mathbb{R}^d$, and let $\{X_k^x\}$ denote the solution of (2.1) with $X_0 = x$. We say that $\{X_k^x: k \geqq 0, x \in K_1\}$ is tight if, given $\varepsilon > 0$, there exists a compact $K_2 \subset \mathbb{R}^d$ such that $P_{0,x}\{X_k \in K_2\} > 1 - \varepsilon$ for all $k \geqq 0$ and $x \in K_1$. Below is our theorem on the convergence of X_k as $k \to \infty$.

THEOREM 1. *Assume that Conditions* 1-3 *hold with* $\alpha > -1$ *and* $\beta > 0$. *Let* $\{X_k\}$ *be given by* (2.1), *and assume that* $\{X_k^x: k \geqq 0, x \in K\}$ *is tight for K a compact set. Then, for* $B/A > C_0$ *and any bounded continuous function* $f(\cdot)$ *on* $\mathbb{R}^d$,

$$\lim_{k\to\infty} E_{0,x}\{f(X_k)\} = \pi(f) \tag{2.4}$$

uniformly for x in a compact set.

Note that since π concentrates on S^*, under the conditions of Theorem 1, we have that $X_k \to S^*$ as $k \to \infty$ in probability.

Theorem 1 is the same as [8, Thm. 2], except there we assumed that (2.2) was valid for all $k \geqq 0$. However, examination of the proof of [8, Thm. 2] shows that we actually established that

$$\lim_{k\to\infty} E_{0,x;k_0,x_0}\{f(X_k)\} = \pi(f) \tag{2.5}$$

uniformly for x_0 in a compact set and all x, only assuming that (2.2) is valid for all $k \geqq k_0$. It is easy to show that (2.4) follows from (2.5) and the assumption that $\{X_k^x: k \geqq 0, x \in K\}$ is tight.

2.2. Tightness. In this section, we consider the tightness of the discrete-time algorithm

$$X_{k+1} = X_k - a_k(\psi_k(X_k) + \eta_k) + b_k\sigma_k(X_k)W_k. \tag{2.6}$$

Here $\{\psi_k(\cdot)\}$ are Borel functions from $\mathbb{R}^d$ to $\mathbb{R}^d$, $\{\sigma_k(\cdot)\}$ are Borel functions from $\mathbb{R}^d$ to $\mathbb{R}$, $\{\eta_k\}$ is a sequence of $\mathbb{R}^d$-valued random variables, and $\{W_k\}, \{a_k\}, \{b_k\}$ are as in § 2.1. Below, we give sufficient conditions for the tightness of $\{X_k^x: k \geqq 0, x \in K\}$, where K is a compact subset of $\mathbb{R}^d$. Note that algorithm (2.6) is somewhat more general than algorithm (2.1). We consider this more general algorithm because it is sometimes convenient to write an algorithm in the form (2.6) (with $\psi_k(x) \neq \nabla U(x)$ for some x, k) to verify tightness, and then to write the algorithm in the form (2.1) to verify convergence. We give an example of this situation when we consider continuous-state Metropolis-type annealing algorithms in §§ 3 and 4.

Let $\mathcal{G}_k = \sigma(X_0, W_0, \cdots, W_{k-1}, \eta_0, \cdots, \eta_{k-1})$. We consider the following conditions ($\alpha, \beta, \gamma_1, \gamma_2$ are constants whose values are specified later).

Condition 4. Let K be a compact subset of $\mathbb{R}^d$. Then

$$\sup_{k;x\in K} |\psi_k(x)| < \infty,$$

$$\overline{\lim_{k,|x|\to\infty}} \frac{|\psi_k(x)|}{|x|} a_k^{\gamma_1} < \infty,$$

$$\varliminf_{k,|x|\to\infty} \frac{|\psi_k(x)|}{|x|} a_k^{\gamma_2} > 0,$$

$$\varliminf_{k,|x|\to\infty} \left\langle \frac{\psi_k(x)}{|\psi_k(x)|}, \frac{x}{|x|} \right\rangle > 0.$$

Condition 5. Let K be a compact subset of $\mathbb{R}^d$. Then

$$\sup_{k;x\in K} |\sigma_k(x)| < \infty, \qquad \varlimsup_{k,|x|\to\infty} \frac{|\sigma_k(x)|}{|x|} < \infty.$$

Condition 6. There exists $L \geqq 0$ such that

$$E\{|\eta_k|^2 | \mathscr{G}_k\} \leqq L a_k^\alpha (|X_k|^2 + 1) \quad \text{w.p.1}, \tag{2.7a}$$

$$|E\{\eta_k | \mathscr{G}_k\}| \leqq L a_k^\beta (|X_k| + 1) \quad \text{w.p.1}. \tag{2.7b}$$

W_k is independent of $\mathscr{G}_k$.

THEOREM 2. *Assume that Conditions* 4–6 *hold with* $\alpha > -1$, $\beta > 0$, *and* $0 \leqq \gamma_2 \leqq \gamma_1 < \frac{1}{2}$. *Let* $\{X_k\}$ *be given by* (2.6), *and let* K *be a compact subset of* $\mathbb{R}^d$. *Then* $\{X_k^x \colon k \geqq 0, x \in K\}$ *is a tight family of random variables.*

Theorem 2 is proved similarly to [8, Thm. 3], where we assumed that $\sigma_k(\cdot) = 1$ and did not allow the bounds in (2.7) to be state-dependent. The extension to the present case is straightforward.

3. Metropolis-type annealing algorithms. In this section, we review the finite-state Metropolis-type Markov-chain annealing algorithm, generalize it to an arbitrary state space, and then specialize it to a class of algorithms for which the results in § 2 can be applied to establish convergence.

The finite-state Metropolis-type annealing algorithm may be described as follows [12]. Assume that the state space Σ is finite set. Let $U(\cdot)$ be a real-valued function on Σ (the "energy" function) and $\{T_k\}$ be a sequence of strictly positive numbers (the "temperature" sequence). Let $q(i,j)$ be a stationary transition probability from i to j, for $i, j \in \Sigma$. The one-step transition probability at time k for the finite-state Metropolis-type annealing chain $\{X_k\}$ is given by

$$\begin{aligned} P\{X_{k+1} = j | X_k = i\} &= q(i,j) s_k(i,j), \quad j \neq i, \\ P\{X_{k+1} = i | X_k = i\} &= 1 - \sum_{j \neq i} q(i,j) s_k(i,j), \end{aligned} \tag{3.1}$$

where

$$s_k(i,j) = \exp\left(-\frac{[U(j) - U(i)]^+}{T_k} \right). \tag{3.2}$$

This nonstationary Markov chain may 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 $s_k(i,j) > \theta_k$, where θ_k is an independent random variable uniformly distributed on the interval $[0, 1]$; otherwise, set $X_{k+1} = i$. Suppose that the stochastic transition matrix $Q = [q(i,j)]$ is symmetric, i.e., $q(i,j) = q(j,i)$, and the temperature T_k is fixed at a constant $T > 0$. Then it is easy to show that the resulting stationary Markov chain has a Gibbs invariant measure with mass $\propto \exp(-U(i)/T)$. Furthermore, if the chain is recurrent, then the chain, in fact, has a unique Gibbs invariant probability measure, and the transition probabilities converge to the Gibbs probabilities as $k \to \infty$ for all initial states. Of course, if a finite-state Markov chain is irreducible, then it is recurrent. There has been much work on the convergence and asymptotic behavior of the nonstationary annealing chain when $T_k \to 0$ [3], [5], [9], [12], [14], [21], [24], [25].

We next generalize the finite-state Metropolis-type annealing algorithm (3.1), (3.2) to a general state space. In the formulation and analysis of general state space Markov chains, it is usually assumed that the state space Σ is a σ-finite measure space, say (Σ, Λ, μ) (see [22, Chap. 1] for a thorough discussion of general state space Markov chains). Let $U(\cdot)$ be a real-valued measurable function on such a Σ, and let $\{T_k\}$ be as above. Let $q(x, y)$ be a stationary transition probability density with respect to μ from x to y, for $x, y \in \Sigma$. The one-step transition probability at time k for the general state Metropolis-type annealing chain $\{X_k\}$ is given by

$$P\{X_{k+1} \in A \mid X_k = x\} = \int_A q(x, y) s_k(x, y)\, d\mu(y) + r_k(x) 1_A(x), \tag{3.3}$$

where

$$s_k(x, y) = \exp\left(-\frac{[U(y) - U(x)]^+}{T_k}\right) \tag{3.4}$$

($r_k(x)$ gives the appropriate normalization, i.e., $r_k(x) = 1 - \int q(x, y) s_k(x, y)\, d\mu(y)$). Note that if μ does not have an atom at x, then $r_k(x)$ is the self transition probability starting at state x at time k. Also, note that (3.3), (3.4) reduces to (3.1), (3.2) when the state space is finite and μ is counting measure. The general state chain may be interpreted (and simulated) similarly to the finite-state chain: here $q(x, y)$ is a conditional probability density for generating a candidate state $\tilde{X}_k = y$, given the current state $X_k = x$. Suppose that the stochastic transition function $Q(x, A) = \int_A q(x, y)\, d\mu(y)$ is symmetric, i.e., $q(x, y) = q(y, x)$, and the temperature T_k is fixed at a constant $T > 0$. Then it is easy to show that the resulting stationary Markov chain has a Gibbs invariant measure with density (with respect to μ) $\propto \exp(-U(x)/T)$. Furthermore, if this measure is finite and the chain is μ-recurrent,[1] then the chain, in fact, has a unique Gibbs invariant probability measure, and the transition probability measure converges to the Gibbs measure (in the total variation norm) as $k \to \infty$ for all initial states [22, Thm. 7.1, p. 30]. It is known that if a chain is μ-irreducible[2] and satisfies a certain condition due to Doeblin [6, Hyp. (D), p. 192], then it is μ-recurrent. In [7, Chap. 3], we use this theory to give some sufficient conditions for the ergodicity of general state Metropolis-type Markov chains when Σ is a compact metric space and μ is a finite Borel measure. However, there has been almost no work on the convergence and asymptotic behavior of the nonstationary annealing chain when $T_k \to 0$, although, when Σ is a compact metric space, we would expect the behavior to be similar to when Σ is finite.

We next specialize the general state Metropolis-type annealing algorithm (3.3), (3.4) to a d-dimensional Euclidean state space. This is the most important case and the one that has seen application [16], [23]. Actually, the Metropolis-type annealing chain that we consider is not exactly a specialization of the general state chain described above. Motivated by our desire to show convergence of the chain by writing it in the

[1] If, for every $x \in \Sigma$ and $A \in \Lambda$ such that $\mu(A) > 0$, $P_{0,x} \bigcup_{k=1}^{\infty} \{X_k \in A\} = 1$, then $\{X_k^x\}$ is μ-recurrent.

[2] If, for every $x \in \Sigma$ and $A \in \Lambda$ such that $\mu(A) > 0$, $P_{0,x} \bigcup_{k=1}^{\infty} \{X_k \in A\} > 0$, then $\{X_k^x\}$ is μ-irreducible.

form of the modified stochastic gradient algorithm (2.1), we are led to choosing a nonstationary Gaussian transition density

$$q_k(x,y)=\frac{1}{(2\pi b_k^2\sigma_k^2(x))^{d/2}}\exp\left(-\frac{1}{2}\frac{|y-x|^2}{b_k^2\sigma_k^2(x)}\right), \tag{3.5}$$

and a state-dependent temperature sequence

$$T_k(x)=\frac{b_k^2\sigma_k^2(x)}{2a_k}\left(=\frac{\text{const}\,\sigma_k^2(x)}{\log\log k}\right), \tag{3.6}$$

where

$$\sigma_k(x)=(\delta_k|x|)\vee 1, \qquad \delta_k\downarrow 0. \tag{3.7}$$

To understand these choices, suppose that x lies in some fixed compact set. Then, for k large enough,

$$q_k(x,y)=\frac{1}{(2\pi b_k^2)^{d/2}}\exp\left(-\frac{1}{2}\frac{|y-x|^2}{b_k^2}\right) \tag{3.8}$$

and

$$T_k(x)=T_k=\frac{b_k^2}{2a_k}. \tag{3.9}$$

The choice of the transition density (3.8) is clear, given that we want to write the chain in the form (2.1). The choice of the temperature schedule (3.9) is also clear if we view (2.1) as a sampled version of the associated diffusion (1.1) with sampling intervals a_k and sampling times $t_k=\sum_{n=0}^{k-1}a_n$, since then we should have the corresponding sampled temperatures $T(t_k)=c^2(t_k)/2$. Indeed, it is straightforward to check that, if $C=B/A$, then

$$T_k=\frac{b_k^2}{2a_k}\sim\frac{c^2(t_k)}{2}=T(t_k)\quad\text{as } k\to\infty$$

(recall that $a_k=A/k$, $b_k^2=B/k\log\log k$, and $c^2(t)=C/\log t$ for large k, t). Finally, the reason that the $|x|$ dependence is needed in $\sigma_k(x)$, and hence both (3.5), (3.6), is that to establish tightness of the annealing chain by writing the chain in the form of (2.6), we need a condition similar to the following:

$$|\psi_k(x)|\geqq\text{const}\,|x|,\qquad |x|\text{ large},\quad k\text{ fixed},$$

for suitable choice of $\psi_k(\cdot)$. In other words, the annealing chain must generate a drift (toward the origin) at least proportional to the distance from the origin. This discussion leads us to the following continuous-state Metropolis-type Markov-chain annealing algorithm and convergence result. To establish convergence, we must assume, along with Conditions 1 and 2, the following condition.

Condition 7. It holds that

$$\inf_{\delta>0}\ \overline{\lim_{|x|\to\infty}}\ \sup_{|y-x|<\delta|x|}|HU(y)|\frac{|x|^2}{U(x)}<\infty.$$

This condition is satisfied if, for example, $U(x)\sim\text{const}\,|x|^p$ and $HU(x)=O(|x|^{p-2})$ as $|x|\to\infty$, for some $p\geqq 2$.

Metropolis-Type Annealing Algorithm 1. Let $\{X_k\}$ be a Markov chain with one-step transition probability at time k given by

$$P\{X_{k+1}\in A\,|\,X_k=x\}=\int_A s_k(x,y)\;dN(x,b_k^2\sigma_k^2(x)I)(y)+r_k(x)1_A(x), \tag{3.10}$$

where

$$\sigma_k(x)=(a_k^\gamma|x|)\vee 1, \tag{3.11}$$

$$s_k(x,y)=\exp\left(-\frac{2a_k}{b_k^2}\frac{[U(y)-U(x)]^+}{\sigma_k^2(x)}\right), \tag{3.12}$$

and $\gamma>0$ ($r_k(x)$ gives the correct normalization).

THEOREM 3. *Assume that Conditions* 1, 2, *and* 7 *hold, and also that*

$$\overline{\lim_{|x|\to\infty}}\frac{|\nabla U(x)|}{|x|}<\infty. \tag{3.13}$$

Let $\{X_k\}$ *be the Markov chain with transition probability given by* (3.10)-(3.12) *and with* $0<\gamma<\frac{1}{4}$. *Then, for* $B/A>C_0$ *and any bounded continuous function* $f(\cdot)$ *on* $\mathbb{R}^d$,

$$\lim_{k\to\infty}E_{0,x}\{f(X_k)\}=\pi(f) \tag{3.14}$$

uniformly for x in a compact set.

The proof of Theorem 3 is in § 4.1. Observe that the conditions of Theorem 3 are satisfied if, for example, $\nabla U(x)\sim\text{const}\,x$ and $HU(x)=O(1)$ as $|x|\to\infty$. We can allow for faster growth in $\nabla U(x)$ by using a suitable modification of (3.12).

Metropolis-Type Annealing Algorithm 2. Let $\{X_k\}$ be a Markov chain with one-step transition probability at time k given by

$$P\{X_{k+1}\in A\,|\,X_k=x\}=\int_A s_k(x,y)\;dN(x,b_k^2\sigma_k^2(x)I)(y)+r_k(x)1_A(x), \tag{3.15}$$

where

$$\sigma_k(x)=(a_k^\gamma|x|)\vee 1, \tag{3.16}$$

$$\begin{aligned} s_k(x,y)&=\exp\left(-\frac{2a_k}{b_k^2}\frac{[U(y)-U(x)]^+}{\sigma_k^2(x)}\right) && \text{if } U(x)\leqq\frac{|x|^2+1}{a_k^\gamma}\\ &=\exp\left(-\frac{2a_k}{b_k^2}\frac{[|y|^2-|x|^2]^+}{\sigma_k^2(x)}\right) && \text{if } U(x)>\frac{|x|^2+1}{a_k^\gamma}, \end{aligned} \tag{3.17}$$

and $\gamma>0$ ($r_k(x)$ gives the correct normalization). Note that if K is any fixed compact, $X_{k=x}\in K$, and k is very large, then (3.17) and (3.12) coincide. Note also that (3.17), like (3.12), only uses measurements of $U(\cdot)$ (and not $\nabla U(\cdot)$).

THEOREM 4. *Assume that Conditions* 1, 2, *and* 7 *hold, and also that*

$$\overline{\lim_{|x|\to\infty}}|\nabla U(x)|\frac{|x|}{U(x)}<\infty. \tag{3.18}$$

Let $\{X_k\}$ *be the Markov chain with transition probability given by* (3.15)-(3.17) *and with* $0<\gamma<\frac{1}{8}$. *Then, for* $B/A>C_0$ *and any bounded continuous function* $f(\cdot)$ *on* $\mathbb{R}^d$,

$$\lim_{k\to\infty}E_{0,x}\{f(X_k)\}=\pi(f) \tag{3.19}$$

uniformly for x in a compact set.

The proof of Theorem 4 is in § 4.2. Observe that the conditions of Theorem 4 are satisfied if, for example, $\nabla U(x)\sim\text{const}\,|x|^{p-2}x$ and $HU(x)=O(|x|^{p-2})$ as $|x|\to\infty$, for some $p\geqq 2$.

4. Proofs of Theorems 3 and 4. In the following, $c_1, c_2, \cdots$ denotes positive constants whose value may change from proof to proof. We need the following lemma.

LEMMA 1. *Assume that* $V(\cdot)$ *is a* C^2 *function from* $\mathbb{R}^d$ *to* $\mathbb{R}$. *Let*

$$s(x, y) = \exp(-\lambda[V(y) - V(x)]^+)$$

and

$$\hat{s}(x, y) = \exp(-\lambda[\langle \nabla V(x), y - x\rangle]^+),$$

where $\lambda > 0$. *Then*

$$|s(x, y) - \hat{s}(x, y)| \leqq \lambda \sup_{\varepsilon \in (0,1)} |HV(x + \varepsilon(y - x))| |y - x|^2$$

for all $x, y \in \mathbb{R}^d$.

Proof. Let

$$f(x, y) = V(y) - V(x) - \langle \nabla V(x), y - x\rangle.$$

Then, by the second-order Taylor theorem,

$$|f(x, y)| \leqq \sup_{\varepsilon \in (0,1)} |HV(x + \varepsilon(y - x))| |y - x|^2. \tag{4.1}$$

By separately considering the four cases corresponding to the possible signs of $V(y) - V(x)$ and $\langle \nabla V(x), y - x\rangle$, it can be shown that

$$|s(x, y) - \hat{s}(x, y)| \leqq 1 - \exp(-\lambda |f(x, y)|) \leqq \lambda |f(x, y)|. \tag{4.2}$$

Combining (4.1) and (4.2) completes the proof. □

4.1. Proof of Theorem 3. We write

$$X_{k+1} = X_k - a_k(\nabla U(X_k) + \xi_k) + b_k W_k \tag{4.3}$$

(this defines ξ_k) and apply Theorem 1 to show that, if $\{X_k^x: k \geqq 0, x \in K\}$ is tight for K compact, then (3.14) is true. We further let $\psi(x) = \nabla U(x)$, write

$$X_{k+1} = X_k - a_k(\psi(X_k) + \eta_k) + b_k \sigma_k(X_k) W_k \tag{4.4}$$

(this defines η_k), and apply Theorem 2 to show that $\{X_k^x: k \geqq 0, x \in K\}$ is, in fact, tight for K compact, and that (3.14) is, in fact, true.

We first show that we can find a version of $\{X_k\}$ in the form

$$X_{k+1} = X_k + b_k \sigma_k(X_k) \zeta_k W_k. \tag{4.5}$$

To do this, we inductively define the sequence $\{W_k, \zeta_k\}$ of random variables as follows. Assume that $X_0, W_0, \cdots, W_{k-1}, \zeta_0, \cdots, \zeta_{k-1}$ have been defined. Let W_k be a standard d-dimensional Gaussian random variable independent of $X_0, W_0, \cdots, W_{k-1}, \zeta_0, \cdots, \zeta_{k-1}$, and let ζ_k be a $\{0, 1\}$-valued random variable with

$$P\{\zeta_k = 1 | X_0, W_0, \cdots, W_k, \zeta_0, \cdots, \zeta_{k-1}\} = s_k(X_k, X_k + b_k \sigma_k(X_k) W_k). \tag{4.6}$$

Using (4.6), it is easy to check that (4.5) is a Markov chain that has transition probability given by (3.10)-(3.12). Hence (4.5) is indeed a version of $\{X_k\}$, and we henceforth always deal with this version.

Now, comparing (4.3) and (4.4) with (4.5), we have that

$$\xi_k = -\nabla U(X_k) + \frac{b_k}{a_k}(1 - \sigma_k(X_k)\zeta_k) W_k \tag{4.7}$$

and

$$\eta_k = -\psi(X_k) + \frac{b_k}{a_k} \sigma_k(X_k)(1 - \zeta_k) W_k. \tag{4.8}$$

Furthermore, it is easy to show that W_k is independent of $\mathcal{F}_k$ and $\mathcal{G}_k$, and also that $P\{\xi_k \in \cdot \,|\, \mathcal{F}_k\} = P\{\xi_k \in \cdot \,|\, X_k\}$ and $P\{\eta_k \in \cdot \,|\, \mathcal{G}_k\} = P(\eta_k \in \cdot \,|\, X_k\}$. We use these facts below.

The following lemmas give the crucial estimates for $E\{|\xi_k|^2|\mathcal{F}_k\}$, $|E\{\xi_k \,|\, \mathcal{F}_k\}|$, $E\{|\eta_k|^2|\mathcal{G}_k\}$, and $|E\{\eta_k \,|\, \mathcal{G}_k\}|$.

LEMMA 2. *Let K be a compact subset of $\mathbb{R}^d$. There exists L, $k_0 \geqq 0$ such that, for every $k \geqq k_0$,*

(a) $|E\{\xi_k \,|\, \mathcal{F}_k\}| \leqq L(a_k/b_k)$ *for all* $X_k \in K$, w.p.1;

(b) $E\{|\xi_k|^2|\mathcal{F}_k\} \leqq L(b_k/a_k)$ *for all* $X_k \in K$, w.p.1.

LEMMA 3. *There exists $L \geqq 0$ such that*

(a) $|E\{\eta_k \,|\, \mathcal{G}_k\}| \leqq L(a_k^{1-2\gamma}/b_k)(|X_k|+1)$ w.p.1;

(b) $E\{|\eta_k|^2|\mathcal{G}_k\} \leqq L(b_k/a_k^{1+\gamma})(|X_k|^2+1)$ w.p.1.

Assume that Lemmas 2 and 3 are true. Then Condition 3 is satisfied with $\alpha = -\frac{1}{2} > -1$ and $0 < \beta < \frac{1}{2}$. Conditions 4–6 are satisfied for $\alpha = -\frac{1}{2} - \gamma > -1$, $0 < \beta < \frac{1}{2} - 2\gamma$, and $\gamma_1 = \gamma_2 = 0$ (recall that we assume that $0 < \gamma < \frac{1}{4}$). Hence Theorems 1 and 2 apply, and Theorem 3 follows. It remains to prove Lemmas 2 and 3. We use the following claim.

CLAIM. *Let $u \in \mathbb{R}^d$ with $|u| = 1$. Then*

(a) $\int_{0 \leqq \langle u,w\rangle \leqq \delta} dN(0,I)(w) = O(\delta)$;

(b) $\int_{0 \leqq \langle u,w\rangle \leqq \delta} w\, dN(0,I)(w) = O(\delta^2)$;

(c) $\int_{0 \leqq \langle u,w\rangle \leqq \delta} w \otimes w\, dN(0,I)(w) = O(\delta)$.

Proof. Let $u_1 = u$, and extend u_1 to an orthonormal basis $\{u_1, \cdots, u_d\}$ for $\mathbb{R}^d$. Then, by changing variables (rotation) and using the mean value theorem, we obtain that

$$\text{(a)} \quad \int_{0 \leqq \langle u,w\rangle \leqq \delta} dN(0,I)(w) = \int_0^\delta \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{v^2}{2}\right) dv = O(\delta);$$

$$\text{(b)} \quad \int_{0 \leqq \langle u,w\rangle \leqq \delta} w\, dN(0,I)(w) = u_1 \int_0^\delta v \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{v^2}{2}\right) dv = O(\delta^2);$$

$$\text{(c)} \quad \int_{0 \leqq \langle u,w\rangle \leqq \delta} w \otimes w\, dN(0,I)(w) = u_1 \otimes u_1 \int_0^\delta v^2 \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{v^2}{2}\right) dv + \sum_{i=2}^d u_i \otimes u_i \int_0^\delta \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{v^2}{2}\right) dv = O(\delta).$$

□

Proof of Lemma 2(a). Since K is compact and $a_k^\gamma \to 0$, we can choose k_0 such that $a_k^\gamma |X_k| \leqq 1$ (and so $\sigma_k(X_k) = 1$) for all $X_k \in K$ and $k \geqq k_0$. Hence, using (4.7) and the fact that $P\{\xi_k \in \cdot \,|\, \mathcal{F}_k\} = P\{\xi_k \in \cdot \,|\, X_k\}$ and W_k is independent of X_k, we have, for $k \geqq k_0$ and $X_k \in K$ (w.p.1), that

$$\begin{aligned} E\{\xi_k \,|\, \mathcal{F}_k\} &= E\{\xi_k \,|\, X_k\} \\ &= -\nabla U(X_k) + \frac{b_k}{a_k} E\{(1-\zeta_k) W_k \,|\, X_k\} \\ &= -\nabla U(X_k) - \frac{b_k}{a_k} E\{W_k E\{\zeta_k \,|\, X_k, W_k\} \,|\, X_k\} \\ &= -\nabla U(X_k) - \frac{b_k}{a_k} E\{W_k P\{\zeta_k = 1 \,|\, X_k, W_k\} \,|\, X_k\} \\ &= -\nabla U(X_k) - \frac{b_k}{a_k} E_{W_k}\{W_k P\{\zeta_k = 1 \,|\, X_k, W_k\}\}. \end{aligned} \tag{4.9}$$

Henceforth, we assume that $k \geqq k_0$ and condition on $X_k = x \in K$. Then, using (4.6), we obtain that

$$E\{\xi_k \mid X_k = x\} = -\nabla U(x) - \frac{b_k}{a_k} \int w s_k(x, x + b_k w)\, dN(0, I)(w). \tag{4.10}$$

Let

$$\hat{s}_k(x, y) = \exp\left(-\frac{2a_k}{b_k^2}[\langle \nabla U(x), y - x\rangle]^+\right) \tag{4.11}$$

and $\tilde{s}_k(x, y) = s_k(x, y) - \hat{s}_k(x, y)$. Using the fact that $HU(\cdot) = O(1)$ on a compact, we obtain that, for any fixed $\delta > 0$,

$$\sup_{\varepsilon \in (0,1)} |HU(x + \varepsilon(y - x))| \leqq c_1,$$

for all $|y - x| < \delta$. Hence, using Lemma 1,

$$|\tilde{s}_k(x, y)| \leqq c_2 \frac{a_k}{b_k^2} |y - x|^2, \qquad |y - x| < \delta. \tag{4.12}$$

Of course,

$$|\tilde{s}_k(x, y)| \leqq 1. \tag{4.13}$$

Using (4.12), (4.13), and a standard estimate for the tail probability of a Gaussian random variable, we obtain, for $i \geqq 0$, that

$$\begin{aligned} &\int |w|^i |\tilde{s}_k(x, x + b_k w)|\, dN(0, I)(w) \\ &\quad \leqq \int_{|w| \leqq \delta / b_k} |w|^i |\tilde{s}_k(x, x + b_k w)|\, dN(0, I)(w) \\ &\qquad + \int_{|w| > \delta / b_k} |w|^i |\tilde{s}_k(x, x + b_k w)|\, dN(0, I)(w) \\ &\quad \leqq c_3 a_k + c_3 \exp\left(-\frac{c_4}{b_k^2}\right) \\ &\quad = O(a_k). \end{aligned} \tag{4.14}$$

Now, expanding (4.10) and using (4.14) gives

$$\begin{aligned} E\{\xi_k \mid X_k = x\} &= -\nabla U(x) - \frac{b_k}{a_k} \int w \hat{s}_k(x, x + b_k w)\, dN(0, I)(w) \\ &\quad - \frac{b_k}{a_k} \int w \tilde{s}_k(x, x + b_k w)\, dN(0, I)(w) \end{aligned}$$

$$\begin{aligned} &= -\nabla U(x) - \frac{b_k}{a_k} \int w \hat{s}_k(x, x + b_k w)\, dN(0, I)(w) \\ &\quad + O(b_k) \end{aligned} \tag{4.15}$$

$$\begin{aligned} &= -\nabla U(x) - \frac{b_k}{a_k} \int_{\langle \nabla U(x), w\rangle \leqq 0} w\, dN(0, I)(w) \\ &\quad - \frac{b_k}{a_k} \int_{\langle \nabla U(x), w\rangle > 0} w \exp\left(-\frac{2a_k}{b_k}\langle \nabla U(x), w\rangle\right) dN(0, I)(w) \\ &\quad + O(b_k). \end{aligned} \tag{4.16}$$

Clearly,

$$E\{\xi_k \mid X_k = x\} = O(b_k) \tag{4.17}$$

for x such that $\nabla U(x) = 0$. Henceforth, we assume that $\nabla U(x) \neq 0$. Let $\nabla \hat{U}(x) = \nabla U(x)/|\nabla U(x)|$. Completing the square in the second integral in (4.16), we obtain that

$$\begin{aligned} E\{\xi_k \mid X_k = x\} = {} & -\nabla U(x) - \frac{b_k}{a_k} \int_{\langle \nabla \hat{U}(x), w\rangle \leq 0} w \, dN(0, I)(w) \\ & - \frac{b_k}{a_k} \int_{\langle \nabla \hat{U}(x), w\rangle \geq 0} w \exp\left(2\left(\frac{a_k}{b_k}\right)^2 |\nabla U(x)|^2\right) dN\left(\frac{2a_k}{b_k} \nabla U(x), I\right)(w) \\ & + O(b_k). \end{aligned} \tag{4.18}$$

Now $\nabla U(x) = O(1)$, and so

$$\exp\left(2\left(\frac{a_k}{b_k}\right)^2 |\nabla U(x)|^2\right) = 1 + O\left(\left(\frac{a_k}{b_k}\right)^2\right). \tag{4.19}$$

Substituting (4.19) into (4.18), using $\nabla U(x) = O(1)$ and $a_k/b_k = O(1)$, and changing variables from $w + 2(a_k/b_k)\nabla U(x)$ to w, gives

$$\begin{aligned} E\{\xi_k \mid X_k = x\} = {} & -\nabla U(x) - \frac{b_k}{a_k} \int_{\langle \nabla \hat{U}(x), w\rangle \leq 0} w \, dN(0, I)(w) \\ & - \frac{b_k}{a_k} \int_{\langle \nabla \hat{U}(x), w\rangle \geq O(a_k/b_k)} w \, dN(0, I)(w) \\ & + 2\nabla U(x) \int_{\langle \nabla \hat{U}(x), w\rangle \geq O(a_k/b_k)} dN(0, I)(w) \\ & + O\left(\frac{a_k}{b_k}\right) + O(b_k) \\ = {} & \frac{b_k}{a_k} \int_{0 \leq \langle \nabla \hat{U}(x), w\rangle \leq O(a_k/b_k)} w \, dN(0, I)(w) \\ & - 2\nabla U(x) \int_{0 \leq \langle \nabla \hat{U}(x), w\rangle \leq O(a_k/b_k)} dN(0, I)(w) \\ & + O\left(\frac{a_k}{b_k}\right). \end{aligned} \tag{4.20}$$

Hence, by (a) and (b) of the claim, and by again using $\nabla U(x) = O(1)$, we have that

$$E\{\xi_k \mid X_k = x\} = O\left(\frac{a_k}{b_k}\right). \tag{4.21}$$

Combining (4.17) and (4.21) completes the proof of Lemma 2(a). □

Proof of Lemma 2(b). As in the proof of Lemma 2(a), choose k_0 such that $a_k^\gamma |X_k| \leq 1$ (and so $\sigma_k(X_k) = 1$) for all $X_k \in K$ and $k \geq k_0$. Hence, using (4.7) and the fact that

$P\{\xi_k \in \cdot \mid \mathscr{F}_k\} = P\{\xi_k \in \cdot \mid X_k\}$, W_k is independent of X_k, and $\nabla U(\cdot) = O(1)$ on a compact, we have, for $k \geqq k_0$ and $X_k \in K$ (w.p.1), that

$$\begin{aligned}
E\{\xi_k \otimes \xi_k \mid \mathscr{F}_k\} &= E\{\xi_k \otimes \xi_k \mid X_k\} \\
&= \left(\frac{b_k}{a_k}\right)^2 E\{((1-\zeta_k)W_k)\otimes((1-\zeta_k)W_k) \mid X_k\} + e_k(X_k) \\
&= \left(\frac{b_k}{a_k}\right)^2 I - \left(\frac{b_k}{a_k}\right)^2 E\{W_k \otimes W_k E\{\zeta_k \mid X_k, W_k\} \mid X_k\} + e_k(X_k) \qquad (4.22)\\
&= \left(\frac{b_k}{a_k}\right)^2 I - \left(\frac{b_k}{a_k}\right)^2 E\{W_k \otimes W_k P\{\zeta_k = 1 \mid X_k, W_k\} \mid X_k\} + e_k(X_k) \\
&= \left(\frac{b_k}{a_k}\right)^2 I - \left(\frac{b_k}{a_k}\right)^2 \underset{W_k}{E}\{W_k \otimes W_k P\{\zeta_k = 1 \mid X_k, W_k\}\} + e_k(X_k),
\end{aligned}$$

where

$$\begin{aligned}
e_k(X_k) &= O\left(\frac{b_k}{a_k}|\nabla U(X_k)| + |\nabla U(X_k)|^2\right) \\
&= O\left(\frac{b_k}{a_k}\right).
\end{aligned}$$

Henceforth, we assume that $k \geqq k_0$ and condition on $X_k = x \in K$. Then, using (4.6), we obtain that

$$\begin{aligned}
E\{\xi_k \otimes \xi_k \mid X_k = x\} = \left(\frac{b_k}{a_k}\right)^2 I - \left(\frac{b_k}{a_k}\right)^2 \int w \otimes w s_k(x, x + b_k w)\, dN(0, I)(w) \\
+ O\left(\frac{b_k}{a_k}\right). \qquad (4.23)
\end{aligned}$$

Let $\hat{s}_k(x, y)$ be given by (4.11) and $\tilde{s}_k(x, y) = s_k(x, y) - \hat{s}_k(x, y)$. Then, expanding (4.23) and using (4.14), gives

$$\begin{aligned}
E\{\xi_k \otimes \xi_k \mid X_k = x\} &= \left(\frac{b_k}{a_k}\right)^2 I - \left(\frac{b_k}{a_k}\right)^2 \int w \otimes w \hat{s}_k(x, x + b_k w)\, dN(0, I)(w) \\
&\quad - \left(\frac{b_k}{a_k}\right)^2 \int w \otimes w \tilde{s}_k(x, x + b_k w)\, dN(0, I)(w) \\
&\quad + O\left(\frac{b_k}{a_k}\right) \\
&= \left(\frac{b_k}{a_k}\right)^2 I - \left(\frac{b_k}{a_k}\right)^2 \int w \otimes w \hat{s}_k(x, x + b_k w)\, dN(0, I)(w) \qquad (4.24)\\
&\quad + O\left(\frac{b_k^2}{a_k}\right) + O\left(\frac{b_k}{a_k}\right) \\
&= \left(\frac{b_k}{a_k}\right)^2 I - \left(\frac{b_k}{a_k}\right)^2 \int_{\langle \nabla U(x), w\rangle \leqq 0} w \otimes w\, dN(0, I)(w) \qquad (4.25)\\
&\quad - \left(\frac{b_k}{a_k}\right)^2 \int_{\langle \nabla U(x), w\rangle > 0} w \otimes w \\
&\quad \cdot \exp\left(-\frac{2a_k}{b_k}\langle \nabla U(x), w\rangle\right) dN(0, I)(w) \\
&\quad + O\left(\frac{b_k}{a_k}\right).
\end{aligned}$$

Clearly,

$$E\{\xi_k \otimes \xi_k \mid X_k = x\} = O\left(\frac{b_k}{a_k}\right) \tag{4.26}$$

for x such that $\nabla U(x) = 0$. Henceforth, we assume that $\nabla U(x) \neq 0$. Let $\nabla \hat{U}(x) = \nabla U(x)/|\nabla U(x)|$. Completing the square in the second integral in (4.25), we obtain that

$$\begin{aligned} E\{\xi_k \otimes \xi_k \mid X_k = x\} &= \left(\frac{b_k}{a_k}\right)^2 I - \left(\frac{b_k}{a_k}\right)^2 \int_{\langle \hat{U}(x), w\rangle \leqq 0} w \otimes w \, dN(0, I)(w) \\ &\quad - \left(\frac{b_k}{a_k}\right)^2 \int_{\langle \nabla \hat{U}(x), w\rangle \geqq 0} w \otimes w \\ &\quad \cdot \exp\left(2\left(\frac{a_k}{b_k}\right)^2 (|\nabla U(x)|)^2\right) dN\left(-\frac{2a_k}{b_k} \nabla U(x), I\right)(w) \\ &\quad + O\left(\frac{b_k}{a_k}\right). \end{aligned} \tag{4.27}$$

Substituting (4.19) into (4.27), using $\nabla U(x) = O(1)$ and $a_k/b_k = O(1)$, and changing variables from $w + 2(a_k/b_k)\nabla U(x)$ to w gives

$$\begin{aligned} E\{\xi_k \otimes \xi_k \mid X_k = x\} &= \left(\frac{b_k}{a_k}\right)^2 I - \left(\frac{b_k}{a_k}\right)^2 \int_{\langle \nabla \hat{U}(x), w\rangle \leqq 0} w \otimes w \, dN(0, I)(w) \\ &\quad - \left(\frac{b_k}{a_k}\right)^2 \int_{\langle \nabla \hat{U}(x), w\rangle \geqq O(a_k/b_k)} w \otimes w \, dN(0, I)(w) \\ &\quad + O(1) + O\left(\frac{b_k}{a_k}\right) \\ &= \left(\frac{b_k}{a_k}\right)^2 \int_{0 \leqq \langle \nabla \hat{U}(x), w\rangle \leqq O(a_k/b_k)} w \otimes w \, dN(0, I)(w) \\ &\quad + O\left(\frac{b_k}{a_k}\right). \end{aligned} \tag{4.28}$$

Hence, by part (c) of the Claim,

$$E\{\xi_k \otimes \xi_k \mid X_k = x\} = O\left(\frac{b_k}{a_k}\right). \tag{4.29}$$

Combining (4.26) and (4.29) and using $|\xi_k|^2 \leqq |\xi_k \otimes \xi_k|$ completes the proof of Lemma 2(b). □

Proof of Lemma 3. Using (4.8) and the fact that $P\{\eta_k \in \cdot \mid \mathscr{G}_k\} = P\{\eta_k \in \cdot \mid X_k\}$, W_k is independent of X_k, and $\psi(x) = \nabla U(x) = O(|x| + 1)$, we get, similarly to (4.9) and (4.22), that

$$E\{\eta_k \mid \mathscr{G}_k\} = -\psi(X_k) - \frac{b_k}{a_k} \sigma_k(X_k) \underset{W_k}{E}\{W_k P\{\zeta_k = 1 \mid X_k, W_k\}\}$$

and

$$E\{\eta_k \otimes \eta_k \mid \mathcal{G}_k\} = \left(\frac{b_k}{a_k}\right)^2 \sigma_k^2(X_k) I$$

$$- \left(\frac{b_k}{a_k}\right)^2 \sigma_k^2(X_k) \underset{W_k}{E}\{W_k \otimes W_k P\{\zeta_k = 1 \mid X_k, W_k\}\} + e_k(X_k),$$

where

$$e_k(X_k) = O\left(\frac{b_k}{a_k} \sigma_k(X_k) \mid \psi(X_k)\mid + \mid\psi(X_k)\mid^2\right)$$

$$= O\left(\frac{b_k}{a_k} \sigma_k(X_k)(|X_k| + 1) + |X_k|^2\right).$$

Henceforth, we condition on $X_k = x$ and assume, for simplicity, that $|x| \geqq 1$ and $a_k \leqq 1$.

Let

$$\hat{s}_k(x, y) = \exp\left(-\frac{2a_k}{b_k^2} \frac{[\langle \nabla U(x), y - x\rangle]^+}{\sigma_k^2(x)}\right) \tag{4.30}$$

and $\tilde{s}_k(x, y) = s_k(x, y) - \hat{s}_k(x, y)$. Using Condition 7, we obtain, for some $\delta > 0$, that

$$\sup_{\varepsilon \in (0,1)} |HU(x + \varepsilon(y - x))| \leqq c_1 \left(\frac{U(x)}{|x|^2} + 1\right)$$

for all $|y - x| < \delta |x|$. By assumption, however, $\nabla U(x) = O(|x|)$, and so, by the mean value theorem, $U(x) = O(|x|^2)$. Hence, using Lemma 1, we obtain that

$$|\tilde{s}_k(x, y)| \leqq c_2 \frac{a_k}{b_k^2} \frac{|y - x|^2}{\sigma_k^2(x)}, \qquad |y - x| < \delta \sigma_k(x),$$

and, similarly to the derivation of (4.14), we obtain that

$$\int |w|^i \tilde{s}_k(x, x + b_k \sigma_k(x) w)\, dN(0, I)(w) = O(a_k). \tag{4.31}$$

Next, using (4.31), we obtain, similarly to the derivation of (4.15) and (4.24), that

$$E\{\eta_k \mid X_k = x\} = -\psi(x) - \frac{b_k}{a_k} \sigma_k(x) \int w \hat{s}_k(x, x + b_k \sigma_k(x) w)\, dN(0, I)(w)$$

$$+ O(b_k \sigma_k(x))$$

and

$$E\{\eta_k \otimes \eta_k \mid X_k = x\} = \left(\frac{b_k}{a_k}\right)^2 \sigma_k^2(x) I$$

$$- \left(\frac{b_k}{a_k}\right)^2 \sigma_k^2(x) \int w \otimes w \hat{s}_k(x, x + b_k \sigma_k(x) w)\, dN(0, I)(w)$$

$$+ O\left(\frac{b_k^2}{a_k} \sigma_k^2(x)\right) + O\left(\frac{b_k}{a_k} \sigma_k(x)|x| + |x|^2\right).$$

At this point, we separately consider the cases where $a_k^\gamma|x|\leqq$ and >1 ($\sigma_k(x)=1$ and $a_k^\gamma|x|$, respectively). Proceeding as in the proof of Lemma 2 and using $\psi(x)=O(|x|)$ and $a_k^{1-\gamma}/b_k=O(1)$, we can show that

$$E\{\eta_k\,|\,X_k=x\}=O\left(\frac{a_k^{1-2\gamma}}{b_k}\right),\qquad a_k^\gamma|x|\leqq 1,$$

$$=O\left(\frac{a_k^{1-\gamma}}{b_k}|x|\right),\qquad a_k^\gamma|x|>1,$$

and

$$E\{\eta_k\otimes\eta_k\,|\,X_k=x\}=O\left(\frac{b_k}{a_k^{1+\gamma}}\right),\qquad a_k^\gamma|x|\leqq 1,$$

$$=O\left(\frac{b_k}{a_k^{1-\gamma}}|x|^2\right),\qquad a_k^\gamma|x|>1.$$

Combining the two cases completes the proof of the lemma. □

Remark 1. In Fig. 1 we demonstrate the type of approximations used in the proof of Theorem 3. In Fig. 1(a) we show the transition density $p_k(x,y)$ for the Markov chain with transition probability given by (3.10)-(3.12); in Figure 1(b) we show the transition density $p_k'(x,y)$ for the same Markov chain but using $\hat{s}_k(x,y)$ (4.30) in place of $s_k(x,y)$ (3.12); and in Fig. 1(c) we show the transition density $p_k''(x,y)$ for the Markov chain of (2.1) with $\xi_k=0$. Note that the densities in Figs. 1(a) and 1(b) contain impulsive components associated with the positive probability of no transition. All three densities are "close" for sufficiently large k and x in a compact set, and this is the basis of the proof. However, for small k, the transition densities can be quite different. In particular, it is seen that the Metropolis-type algorithm takes a less "local" point of view than the gradient-based algorithms.

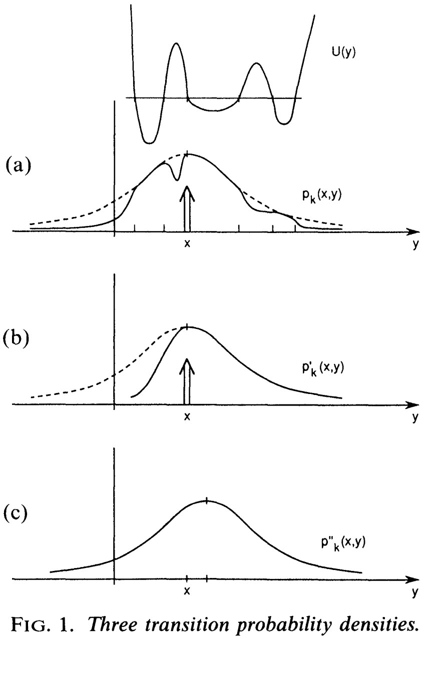

FIG. 1. *Three transition probability densities.*

Remark 2. The Metropolis-type Markov-chain annealing algorithms use only measurements of $U(\cdot)$ (and not $\nabla U(\cdot)$). Another class of algorithms that use only measurements of $U(\cdot)$ could be based on a finite-difference approximation $D_kU(\cdot)$ of $\nabla U(\cdot)$,

$$X_{k+1}=X_k-a_kD_kU(X_k)+b_kW_k. \tag{4.32}$$

Suppose that $D_kU(\cdot)$ is a random direction forward finite-difference approximation; i.e., suppose that θ_k is an independent random vector uniformly distributed on the d-one-dimensional unit sphere, and

$$D_kU(x)=\frac{U(x+h_k\theta_k)-U(x)}{h_k}\,\theta_k$$

($\{h_k\}$ is a sequence of nonzero numbers with $h_k\to 0$). If we write (4.32) in the form (2.1), then, by analysis similar to [19, pp. 58-60], it can be shown that ξ_k is *bounded* for X_k in a compact. However, when we write (3.10)-(3.12) in the form (2.1), the best estimate we can obtain suggests that ξ_k is *unbounded* for X_k in a compact (see Lemma 2(b), and note that $b_k/a_k\to\infty$). Hence the Metropolis-type approximation appears to be much farther away from an exact gradient-based algorithm than a finite-difference approximation.

4.2. Proof of Theorem 4. We write

$$X_{k+1}=X_k-a_k(\nabla U(X_k)+\xi_k)+b_kW_k$$

(this defines ξ_k) and apply Theorem 1 to show that, if $\{X_k^x\colon k\geqq 0, x\in K\}$ is tight for K compact, then (3.19) is true. We further let

$$\begin{aligned}\psi_k(x)&=\nabla U(x)\quad \text{if } U(x)\leqq\frac{|x|^2+1}{a_k^\gamma}\\ &=2x\qquad\quad \text{if } U(x)>\frac{|x|^2+1}{a_k^\gamma},\end{aligned}$$

write

$$X_{k+1}=X_k-a_k(\psi_k(X_k)+\eta_k)+b_k\sigma_k(X_k)W_k$$

(this defines η_k), and apply Theorem 2 to show that $\{X_k^x\colon k\geqq 0, x\in K\}$ is, in fact, tight for K compact and (3.19) is, in fact, true.

The following lemmas give the crucial estimates for $E\{|\xi_k|^2|\mathcal{F}_k\}$, $|E\{\xi_k\,|\,\mathcal{F}_k\}|$, $E\{|\eta_k|^2|\mathcal{G}_k\}$, and $|E\{\eta_k\,|\,\mathcal{G}_k\}|$ (compare with Lemmas 2 and 3).

LEMMA 4. *Let K be a compact subset of $\mathbb{R}^d$. Then there exists $L, k_0\geqq 0$ such that, for every $k\geqq k_0$,*

(a) $|E\{\xi_k\,|\,\mathcal{F}_k\}|\leqq L(a_k/b_k)$ *for all* $X_k\in K$, w.p.1;

(b) $E\{|\xi_k|^2|\mathcal{F}_k\}\leqq L(b_k/a_k)$ *for all* $X_k\in K$, w.p.1.

LEMMA 5. *There exists $L\geqq 0$ such that*

(a) $|E\{\eta_k\,|\,\mathcal{G}_k\}|\leqq L(a_k^{1-4\gamma}/b_k)(|X_k|+1)$ w.p.1;

(b) $E\{|\eta_k|^2|\mathcal{G}_k\}\leqq L(b_k/a_k^{1+2\gamma})(|X_k|^2+1)$ w.p.1.

Assume that Lemmas 4 and 5 are true. Then Condition 3 is satisfied with $\alpha=-\frac{1}{2}>-1$ and $0<\beta<\frac{1}{2}$. Conditions 4-6 are satisfied with $\alpha=-\frac{1}{2}-2\gamma>-1, 0<\beta<\frac{1}{2}-4\gamma$, $\gamma_1=\gamma$, and $\gamma_2=0$ (recall that we assume that $0<\gamma<\frac{1}{8}$). We note that the second relation in Condition 4 is verified with $\gamma_1=\gamma$ by considering the two cases where $U(x)$ is $<$ or $\geqq$ $(|x|^2+1)/a_k^\gamma$ and applying (3.18); in fact, we obtain that $\psi_k(x)=O(|x|/a_k^\gamma)$ as $|x|\to\infty$ uniformly for all k. Hence Theorems 1 and 2 apply, and Theorem 4 follows. It remains to prove Lemmas 4 and 5.

Proof of Lemma 4. Since K is compact and $a_k^\gamma \to 0$, we can choose k_0 such that $U(X_k) \leqq (|X_k|^2+1)/a_k^\gamma$ for all $X_k \in K$ and $k \geqq k_0$. Hence the proof of Lemma 4 is the same as that of Lemma 2. □

Proof of Lemma 5. Using $\psi_k(x) = O(|x|/a_k^\gamma + 1)$ (see discussion following the statement of Lemmas 4 and 5), we get, similarly to the proof of Lemma 3, that

$$E\{\eta_k \mid \mathscr{G}_k\} = -\psi_k(X_k) - \frac{b_k}{a_k}\sigma_k(X_k) \underset{W_k}{E}\{W_k P\{\zeta_k = 1 \mid X_k, W_k\}\}$$

and

$$\begin{aligned} E\{\eta_k \otimes \eta_k \mid \mathscr{G}_k\} = {} & \left(\frac{b_k}{a_k}\right)^2 \sigma_k^2(X_k) I \\ & - \left(\frac{b_k}{a_k}\right)^2 \sigma_k^2(X_k) \underset{W_k}{E}\{W_k \otimes W_k P\{\zeta_k = 1 \mid X_k, W_k\}\} + e_k(X_k), \end{aligned}$$

where

$$\begin{aligned} e_k(X_k) &= O\left(\frac{b_k}{a_k}\sigma_k(X_k)|\psi_k(X_k)| + |\psi_k(X_k)|^2\right) \\ &= O\left(\frac{b_k}{a_k}\sigma_k(X_k)\left(\frac{|X_k|}{a_k^\gamma} + 1\right) + \frac{|X_k|^2}{a_k^{2\gamma}}\right). \end{aligned}$$

Henceforth, we condition on $X_k = x$ and assume for simplicity that $|x| \geqq 1$ and $a_k \leqq 1$.

Let

$$\begin{aligned} \hat{s}_k(x, y) &= \exp\left(-\frac{2a_k}{b_k^2}\frac{[\langle \nabla U(x), y - x\rangle]^+}{\sigma_k^2(x)}\right) \quad \text{if } U(x) \leqq \frac{|x|^2+1}{a_k^\gamma} \\ &= \exp\left(-\frac{2a_k}{b_k^2}\frac{[\langle 2x, y - x\rangle]^+}{\sigma_k^2(x)}\right) \quad \text{if } U(x) > \frac{|x|^2+1}{a_k^\gamma} \end{aligned}$$

and $\tilde{s}_k(x, y) = s_k(x, y) - \hat{s}_k(x, y)$. Now if a C^2 function $V(\cdot)$ satisfies Condition 7, then, for some $\delta > 0$,

$$\sup_{\varepsilon \in (0,1)} |HV(x + \varepsilon(y - x))| \leqq c_1\left(\frac{V(x)}{|x|^2} + 1\right)$$

for all $|y - x| < \delta|x|$; so this inequality holds when $V(z) = U(z)$ and when $V(z) = |z|^2$. Hence, by considering the two cases when $U(x)$ is $\leqq$ or $>$ $(|x|^2+1)/a_k^\gamma$ and using Lemma 1, we obtain that

$$|\tilde{s}_k(x, y)| \leqq c_2 \frac{a_k^{1-\gamma}}{b_k^2}\frac{|y - x|^2}{\sigma_k^2(x)}, \qquad |y - x| < \delta\sigma_k(x),$$

and, similarly to the derivation of (4.14), we obtain that

$$\int |w|^i \tilde{s}_k(x, x + b_k\sigma_k(x)w)\, dN(0, I)(w) = O(a_k^{1-\gamma}). \tag{4.33}$$

Next, using (4.33), we obtain, similarly to the derivation of (4.15) and (4.24), that

$$\begin{aligned} E\{\eta_k \mid X_k = x\} = {} & -\psi_k(x) - \frac{b_k}{a_k}\sigma_k(x) \int w\hat{s}_k(x, x + b_k\sigma_k(x)w)\, dN(0, I)(w) \\ & + O\left(\frac{b_k}{a_k^\gamma}\sigma_k(x)\right) \end{aligned}$$

and

$$E\{\eta_k \otimes \eta_k \mid X_k = x\} = \left(\frac{b_k}{a_k}\right)^2 \sigma_k^2(x) I$$
$$-\left(\frac{b_k}{a_k}\right)^2 \sigma_k^2(x) \int w \otimes w \hat{s}_k(x, x + b_k\sigma_k(x)w)\, dN(0, I)(w)$$
$$+ O\left(\frac{b_k^2}{a_k^{1+\gamma}} \sigma_k^2(x)\right) + O\left(\frac{b_k}{a_k^{1+\gamma}} \sigma_k(x)|x| + \frac{|x|^2}{a_k^{2\gamma}}\right).$$

We now separately consider the cases where $a_k^\gamma |x| \leqq$ and >1 ($\sigma_k(x) = 1$ and $a_k^\gamma |x|$, respectively). Proceeding as in the proof of Lemma 2 and using $\psi_k(x) = O(|x|/a_k^\gamma)$ and $a_k^{1-2\gamma}/b_k = O(1)$, we can show that

$$E\{\eta_k \mid X_k = x\} = O\left(\frac{a_k^{1-4\gamma}}{b_k}\right), \qquad a_k^\gamma |x| \leqq 1,$$
$$= O\left(\frac{a_k^{1-3\gamma}}{b_k} |x|\right), \qquad a_k^\gamma |x| > 1$$

and

$$E\{\eta_k \otimes \eta_k \mid X_k = x\} = O\left(\frac{b_k}{a_k^{1+2\gamma}}\right), \qquad a_k^\gamma |x| \leqq 1,$$
$$= O\left(\frac{b_k}{a_k} |x|^2\right), \qquad a_k^\gamma |x| > 1.$$

Combining the two cases completes the proof of the lemma. □

Acknowledgments. The authors thank the referees for a careful reading of the manuscript, which uncovered an important technical problem, and for suggesting its solution.

REFERENCES

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

[2] V. CERNY, *A thermodynamical approach to the travelling salesman problem*, J. Optim. Theory Appl., 45 (1985), pp. 41-51.

[3] T. S. CHIANG AND Y. CHOW, *On the convergence rate of annealing processes*, SIAM J. Control Optim., 26 (1988), pp. 1455-1470.

[4] T. S. CHIANG, C. R. HWANG, AND S. J. SHEU, *Diffusion for global optimization in* $\mathbb{R}^n$, SIAM J. Control Optim., 25 (1987), pp. 737-752.

[5] D. P. CONNORS AND P. R. KUMAR, *Simulated annealing type Markov chains and their order balance equations*, SIAM J. Control Optim., 27 (1989), pp. 1440-1461.

[6] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1953.

[7] S. B. GELFAND, *Analysis of simulated annealing type algorithms*, Ph.D. thesis, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Report No. LIDS-TH-1668, Cambridge, MA, 1987.

[8] S. B. GELFAND AND S. K. MITTER, *Recursive stochastic algorithms for global optimization in* $\mathbb{R}^d$, SIAM J. Control Optim., 29 (1991), pp. 999-1018.

[9] S. GEMAN AND D. GEMAN, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Trans. Pattern Anal. Machine Intelligence, PAMI-6 (1984), pp. 721-741.

[10] S. GEMAN AND C. R. HWANG, *Diffusions for global optimization*, SIAM J. Control Optim., 24 (1986), pp. 1031-1043.

[11] B. GIDAS, *Global optimization via the Langevin equation*, in Proc. IEEE Conf. on Decision and Control, Fort Lauderdale, FL, 1985, pp. 774-778.

[12] ———, *Nonstationary Markov chains and convergence of the annealing algorithm*, J. Statist. Phys., 39 (1985), pp. 73–131.

[13] U. GRENENDER, *Tutorial in Pattern Theory*, Division of Applied Mathematics, Brown University, Providence, RI, 1984.

[14] B. HAJEK, *Cooling schedules for optimal annealing*, Math. Oper. Res., 13 (1988), pp. 311–329.

[15] C. R. HWANG, *Laplaces method revisited: Weak convergence of probability measures*, Ann. Probab., 8 (1980), pp. 1177–1182.

[16] F. C. JENG AND J. W. WOODS, *Simulated annealing in compound Gaussian random fields*, IEEE Trans. Inform. Theory, IT-36 (1990), pp. 94–107.

[17] S. KIRKPATRICK, C. D. GELATT, AND M. VECCHI, *Optimization by simulated annealing*, Science, 220 (1983), pp. 621–680.

[18] H. J. KUSHNER, *Asymptotic global behavior for stochastic approximation and diffusions with slowly decreasing noise effects: Global minimization via Monte Carlo*, SIAM J. Appl. Math., 47 (1987), pp. 169–185.

[19] H. J. KUSHNER AND D. CLARK, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer, Berlin, 1978.

[20] L. LJUNG, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 551–575.

[21] D. MITRA, F. ROMEO, AND A. SANGIOVANNI-VINCENTELLI, *Convergence and finite-time behavior of simulated annealing*, Adv. Appl. Probab., 18 (1986), pp. 747–771.

[22] S. OREY, *Limit Theorems for Markov Chain Transition Probabilities*, Van Nostrand Reinhold, London, 1971.

[23] T. SIMCHONY, R. CHELLAPA, AND Z. LICHTENSTEIN, *Relaxation algorithms for* MAP *estimation of gray-level images with multiplicative noise*, IEEE Trans. Inform. Theory, IT-36 (1990), pp. 608–614.

[24] J. TSITSIKLIS, *A survey of large time asymptotics of simulated annealing algorithms*, Report No. LIDS-P-1623, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, 1986.

[25] ———, *Markov chains with rare transitions and simulated annealing*, Math. Oper. Res., 14 (1989), pp. 70–90.