JOURNAL OF OPTIMIZATION THEORY AND APPLICATIONS: Vol. 68, No. 3, MARCH 1991

Weak Convergence of Markov Chain Sampling Methods and Annealing Algorithms to Diffusions[1]

S. B. GELFAND[2] AND S. K. MITTER[3]

Communicated by R. Conti

Abstract. Simulated annealing algorithms have traditionally been developed and analyzed along two distinct lines: Metropolis-type Markov chain algorithms and Langevin-type Markov diffusion algorithms. Here, we analyze the dynamics of continuous state Markov chains which arise from a particular implementation of the Metropolis and heat-bath Markov chain sampling methods. It is shown that certain continuous-time interpolations of the Metropolis and heat-bath chains converge weakly to Langevin diffusions running at different time scales. This exposes a close and potentially useful relationship between the Markov chain and diffusion versions of simulated annealing.

Key Words. Annealing algorithms, sampling methods, diffusion approximation, global optimization.

1. Introduction

Simulated annealing algorithms have classically been developed along two distinct lines. Initially, a simulated annealing algorithm for discrete (combinatorial) optimization was suggested in Ref. 1 and was based on simulating a Metropolis-type Markov chain. Later, a simulated annealing algorithm for continuous (multivariate) optimization was suggested in Ref. 2 and was based on simulating a Langevin-type Markov diffusion. The idea behind both of these algorithms is to simulate an imaginary physical system at or near thermal equilibrium whose energy function is identified with the

[1] The research reported here has been supported by the Whirlpool Foundation, by the Air Force Office of Scientific Research under Contract 89-0276, and by the Army Research Office under Contract DAAL-03-86-K-0171 (Center for Intelligent Control Systems).

[2] Assistant Professor, School of Electrical Engineering, Purdue University, West Lafayette, Indiana.

[3] Professor, Department of Electrical Engineering and Computer Science and Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts.

cost function to be minimized. The temperature of the system is slowly decreased to zero and the system is cooled or "annealed" into low energy states.

Following Gidas (Ref. 3), we shall refer to the optimization algorithm based on simulating a Metropolis-type Markov chain as the annealing algorithm and to the optimization algorithm based on simulating a Langevin-type Markov diffusion as the Langevin algorithm. Both the annealing and Langevin algorithms have been applied to a variety of problems and have been the subject of a large amount of theoretical analysis, some of which have generated fundamentally new results about the asymptotic behavior of certain classes of nonstationary Markov chains and diffusions. See Ref. 4 for a guide to the literature. Although the discrete-state annealing algorithm has been the focus of much of the literature, it has also been suggested that a continuous-state annealing algorithm might be effective for certain continuous optimization problems, and some supporting numerical work has been done (Ref. 5). However, we are not aware of any theoretical analysis for such an algorithm, and the analysis of the continuous-state case does not follow from the discrete-state case in a straightforward way.

In this paper, 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 (Ref. 6). 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 gives a precise connection between what is often viewed as artificial stochastic dynamics and a more familiar stochastic dynamics for, say, a particle in a viscous fluid. 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 are valid for both fixed-temperature sampling methods and decreasing-temperature annealing algorithms. Hence, this work exposes a close relationship between the annealing and the Langevin algorithms, other than the fact that both are Markov processes which have a Gibbs invariant distribution for a fixed value of the temperature parameter. Such a relationship provides an important step toward the analysis of the continuous-state annealing algorithm. Indeed, a first step in the analysis of the asymptotic, large-time behavior of a large class of discrete-time recursive stochastic algorithms is to show weak convergence to a continuous-time limit (Refs. 7, 8).

The paper is organized as follows. In Section 2, we describe various Markov chain sampling methods and annealing algorithms, and then state a theorem regarding the weak convergence of these processes and discuss its implications. In Section 3, we prove the theorem using a result of Kushner (Ref. 9).

2. Main Results and Discussion

We first deal with the weak convergence of fixed-temperature Markov chain sampling methods to Langevin diffusions, and then indicate the extension to the weak convergence of decreasing-temperature annealing algorithms to the Langevin algorithm.

We start by reviewing the discrete-state Metropolis and heat-bath Markov chain sampling methods (Ref. 6). Assume that the state space Σ is countable. Let $U(\cdot)$ be a real-valued function on Σ, the energy function for the system under consideration. Also, let T be the strictly positive absolute temperature of the system, and let k_B denote the Boltzmann constant. Let q_{ij} be a stationary transition probability from i to j for $i, j \in \Sigma$. The transition probability from i to j for the Metropolis Markov chain is given by

$$p_{ij} = q_{ij}, \qquad \text{if } U(j) \leq U(i), \tag{1a}$$

$$p_{ij} = q_{ij} \exp[-(U(j) - U(i))/k_B T], \qquad \text{if } U(j) > U(i), \tag{1b}$$

for $i, j \in \Sigma$ with $i \neq j$. The transition probability from i to j for the heat-bath Markov chain is given by

$$p_{ij} = q_{ij} \frac{\exp[-(U(j) - U(i))/k_B T]}{1 + \exp[-(U(j) - U(i))/k_B T]}, \tag{2}$$

for $i, j \in \Sigma$ with $i \neq j$. In both methods, p_{ii} is chosen to give the proper normalization, i.e.,

$$p_{ii} = 1 - \sum_{j \neq i} p_{ij}.$$

Let

$$\pi_i = (1/Z) \exp(-U(i)/k_B T), \qquad i \in \Sigma,$$
$$Z = \sum_i \exp(-U(i)/k_B T);$$

assume $Z < \infty$. If the stochastic matrix $Q = [q_{ij}]$ is symmetric and irreducible, then the detailed balance equation

$$\pi_i p_{ij} = \pi_j p_{ji}, \qquad i, j \in \Sigma,$$

is satisfied, and it follows easily that π_i, $i \in \Sigma$, are the unique stationary probabilities for either the Metropolis or heat-bath Markov chains. If we let $\{X_k\}$ denote either of these chains, and let X be a random variable with $P\{X=i\}=\pi_i$ for $i\in\Sigma$, then $X_k\to X$ in distribution as $k\to\infty$; and, for any bounded Borel function $f(\cdot)$ on Σ,

$$(1/k)\sum_{1}^{k} f(X_n)\to E\{f(X)\}, \qquad \text{w.p. 1, as } k\to\infty;$$

see Ref. 10. Hence, the Metropolis of heat-bath chains may be used to sample from and to compute mean values of functionals with respect to a Gibbs distribution. The Metropolis or heat-bath chains can be interpreted and simulated in the following manner. Write $p_{ij}=q_{ij}s_{ij}+\gamma_i\delta_{ij}$ where δ_{ij} is the Kronecker-delta function. Given the current state $X_k=i$, generate a candidate state $\tilde{X}_k=j$ with probability q_{ij}. Set the next state $X_{k+1}=j$, if $s_{ij}>\Theta_k$, where Θ_k is an independent random variable uniformly distributed on the interval $[0, 1]$; otherwise, set $X_{k+1}=i$.

We next generalize the discrete state Markov chains sampling methods described above to a continuous d-dimensional Euclidean state space. Henceforth, we shall use boldface for vectors and matrices, and subscripts for their components, e.g., x_i will be the ith component of a vector $\mathbf{x}\in\mathbb{R}^d$, and a_{ij} will be the (i, j)th component of a matrix $\mathbf{a}\in\mathbb{R}^{d\times e}$. Let $U(\cdot)$ be a smooth real-valued function on $\mathbb{R}^d$ (we shall make more precise assumptions on $U(\cdot)$ in the sequel). Let $q(\mathbf{x},\mathbf{y})$ be a stationary transition density from $\mathbf{x}$ to $\mathbf{y}$ for $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$. Let

$$s_{\mathrm{M}}(\mathbf{x},\mathbf{y})=1, \qquad \text{if } U(\mathbf{y})\le U(\mathbf{x}), \tag{3a}$$

$$s_{\mathrm{M}}(\mathbf{x},\mathbf{y})=\exp[-(U(\mathbf{y})-U(\mathbf{x}))/k_{\mathrm{B}}T], \qquad \text{if } U(\mathbf{y})> U(\mathbf{x}), \tag{3b}$$

$$s_{\mathrm{H}}(\mathbf{x},\mathbf{y})=\frac{\exp[-(U(\mathbf{y})-U(\mathbf{x}))/\mathrm{k}_{\mathrm{B}}T]}{1+\exp[-(U(\mathbf{y})-U(\mathbf{x}))/k_{\mathrm{B}}T]}, \tag{4}$$

for all $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$. Now, let $\{X_k\}$ be an $\mathbb{R}^d$-valued Markov chain with transition density from $\mathbf{x}$ to $\mathbf{y}$ given by

$$p(\mathbf{x},\mathbf{y})=q(\mathbf{x},\mathbf{y})s(\mathbf{x},\mathbf{y})+\gamma(\mathbf{x})\delta(\mathbf{y}-\mathbf{x}), \tag{5}$$

where

$$\gamma(\mathbf{x})=1-\int q(\mathbf{x},\mathbf{y})s(\mathbf{x},\mathbf{y})\,d\mathbf{y}$$

and $\delta(\cdot)$ is the Dirac-delta function. Here, $s(\cdot,\cdot)=s_{\mathrm{M}}(\cdot,\cdot)$ and $s(\cdot,\cdot)=s_{\mathrm{H}}(\cdot,\cdot)$ for the generalized Metropolis and heat-bath chains, respectively. Note that, if $q(\mathbf{x},\cdot)$ has no impulse at $\mathbf{x}$, then $\gamma(\mathbf{x})$ is the self-transition probability starting at state $\mathbf{x}$. Also note that (5) reduces to (1), (2) when the state space is discrete.

The continuous-state Metropolis and heat-bath Markov chains can be interpreted and simulated analogously to the discrete-state versions. In particular, $q(\mathbf{x}, \mathbf{y})$ is a conditional probability density for generating a candidate state $\tilde{X}_k = \mathbf{y}$, given the current state $X_k = \mathbf{x}$. 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 depends only on the current value of the selected component. Let $r(x_i, y_i)$ be a transition density from x_i to y_i for $x_i, y_i \in \mathbb{R}$. Then, we set

$$\begin{aligned} p(\mathbf{x}, \mathbf{y}) &= (1/d) \sum_{i=1}^{d} s(\mathbf{x}, \mathbf{y}) r(x_i, y_i) \prod_{j \neq i} \delta(y_j - x_j) + \gamma(\mathbf{x}) \delta(\mathbf{y} - \mathbf{x}) \\ &= (1/d) \sum_{i=1}^{d} s(i, \mathbf{x}, y_i) r(x_i, y_i) \prod_{j \neq i} \delta(y_j - x_j) + \gamma(\mathbf{x}) \delta(\mathbf{y} - \mathbf{x}), \end{aligned} \tag{6}$$

where

$$s(i, \mathbf{x}, y_i) = s((x_1, \ldots, x_d), (x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_d)), \tag{7}$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $i = 1, \ldots, d$. Here, we have used the fact that $s(\mathbf{x}, \cdot)$ is bounded and continuous for each $\mathbf{x}$.

Suppose that we take

$$r(x_i, y_i) = 1(x_i = -1)\delta(y_i - 1) + 1(x_i = 1)\delta(y_i + 1), \qquad x_i, y_i \in \mathbb{R},$$

where $1(A)$ is the indicator of the expression A. In this case, if the ith 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 $X_{k,i}$ is ∓ 1. If, in addition,

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

then $\{X_k\}$ corresponds to a discrete-time kinetic Ising model with interaction energies J_{ij} and no external field (Ref. 6).

Suppose instead that we take

$$r(x_i, y_i) = (1/\sqrt{2\pi\sigma^2}) \exp[-(y_i - x_i)^2/2\sigma^2], \qquad x_i, y_i \in \mathbb{R}. \tag{8}$$

In this case, if the ith 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 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 $\epsilon > 0$, let $r_\epsilon(\cdot, \cdot)$ denote the transition density in (8) with $\sigma^2 = \epsilon$, and let $p_\epsilon(\cdot, \cdot)$ denote the corresponding transition density in (6). Let $\{X_k^\epsilon\}$ denote the Markov chain with transition density $p_\epsilon(\cdot, \cdot)$ and initial condition $X_0^\epsilon = X_0$. Interpolate $\{X_k^\epsilon\}$ into a continuous-time process $\{\mathbf{x}^\epsilon(t), t \geq 0\}$ by setting

$$\mathbf{x}^\epsilon(t) = X^\epsilon_{\lfloor t/\epsilon \rfloor}, \qquad t \geq 0,$$

where $\lfloor \alpha \rfloor$ is the largest integer less than or equal to α. Let $D^d[0, \infty)$ denote the space of $\mathbb{R}^d$-valued functions on $[0, \infty)$ which are right-continuous on $[0, \infty)$ and have left-hand limits on $(0, \infty)$, with the Skorohod topology (see Ref. 11). Obviously, $\mathbf{x}^\epsilon(\cdot)$ takes values in $D^d[0, \infty)$. To establish the weak convergence of $\mathbf{x}^\epsilon(\cdot)$ as $\epsilon \to 0$, we will require the following condition on $U(\cdot)$:

(A) $U(\cdot)$ is continuously differentiable, and $U_x(\cdot)$ is bounded and Lipschitz continuous.

Here is our main result.

Theorem 2.1. Assume (A). Then, there is a standard d-dimensional Wiener process $\mathbf{w}(\cdot)$ and a process $\mathbf{x}(\cdot)$, nonanticipative with respect to $\mathbf{w}(\cdot)$, such that $\mathbf{x}^\epsilon(\cdot) \to \mathbf{x}(\cdot)$ weakly in $D^d[0, \infty)$ as $\epsilon \to 0$ and the following results hold:

(a) for the Metropolis method,

$$d\mathbf{x}(t) = -[U_x(\mathbf{x}(t))/2k_B T]\, dt + d\mathbf{w}(t), \qquad t \geq 0, \tag{9}$$

with $\mathbf{x}(0) = X_0$ in distribution;

(b) for the heat-bath method,

$$d\mathbf{x}(t) = -[U_x(\mathbf{x}(t))/4k_B T]\, dt + (1/\sqrt{2})\, d\mathbf{w}(t), \qquad t \geq 0, \tag{10}$$

with $\mathbf{x}(0) = X_0$ in distribution.

The proof of Theorem 2.1 is carried out in Section 3.

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 that $\mathbf{y}(\cdot)$ is a solution of the Langevin equation

$$d\mathbf{y}(t) = -U_y(\mathbf{y}(t))\, dt + \sqrt{2k_B T}\, d\mathbf{w}(t), \qquad t \geq 0,$$

with $\mathbf{y}(0) = X_0$ in distribution. Then, for $\tau(t) = t/2k_B T$, $\mathbf{y}(\tau(\cdot))$ has the same multivariate distributions as $\mathbf{x}(\cdot)$ satisfying (9), while for $\tau(t) = t/4k_B T$, $\mathbf{y}(\tau(\cdot))$ has the same multivariate distributions as $\mathbf{x}(\cdot)$ satisfying (10). Observe that the limit diffusion for the Metropolis chain runs at twice the rate of the limit diffusion for the heat-bath chain, independent of the temperature.

To obtain discrete-state annealing algorithms, we simply replace the fixed temperature T in the discrete-state Markov chain sampling methods by a temperature schedule $\{T_k\}$, where typically $T_k \to 0$ as $k \to \infty$. The resulting Markov chains are nonstationary with one-step transition probabilities $p_{ij}(k)$ given by the r.h.s. of (1) and (2) with T replaced by T_k. Under suitable condition on $\{T_k\}$, $U(\cdot)$, and $\{q_{ij}\}$, it can be shown that $X_k \to S$ in probability as $k \to \infty$, where S is the set of global minima of $U(\cdot)$ (Ref. 12).

To obtain continuous-state annealing algorithms we similarly replace the fixed temperature T in the continuous-state Markov chain sampling methods by a temperature schedule $\{T_k\}$. We are not aware of any analysis concerning annealing algorithms of this type. Suppose that $T(\cdot)$ is a positive continuous function on $[0, \infty)$, where typically $T(t) \to 0$ as $t \to \infty$. For $\epsilon > 0$, let

$$T_k^\epsilon = T(k\epsilon), \qquad k = 0, 1, \ldots,$$

and let $\{X_k^\epsilon\}$ now denote the continuous-state annealing chain with temperature schedule $\{T_k^\epsilon\}$. By a slightly modified argument (see the proof in Section 3), it can be shown that Theorem 2.1 is valid with T replaced by $T(t)$ in (9) and (10). Hence, these annealing algorithms converge weakly to a time-scaled version of the Langevin algorithm

$$d\mathbf{y}(t) = -U_{\mathbf{y}}(\mathbf{y}(t)\, dt + \sqrt{2k_B T(t)}\, d\mathbf{w}(t).$$

Under suitable conditions on $T(\cdot)$ and $U(\cdot)$, it can be shown that $\mathbf{y}(t) \to S$ in probability as $t \to \infty$, where S is the set of global minima of $U(\cdot)$ (Ref. 13).

The weak convergence of a suitably scaled annealing algorithm to the Langevin algorithm potentially provides a great deal of information about the behavior of the annealing algorithm in terms of the corresponding behavior of the Langevin algorithm, which is much easier to analyze. However, this weak convergence and the convergence of the Langevin algorithm in probability to the globally minimum energy states does not directly imply the convergence of the annealing algorithm to the globally minimum energy states; further conditions are required. See Ref. 8 for a discussion of these issues. However, establishing the weak convergence is an important first step in this regard. 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 (the annealing algorithm) weakly converging to a nonstationary SDE limit (the Langevin algorithm). More work needs to be done on this point. Some related work on the convergence of discrete-time recursive stochastic gradient algorithms can be found in Ref. 14.

3. Proof of Theorem 2.1

In this section, we prove Theorem 2.1. We shall make use of the following result of Kushner (Ref. 9) on the weak convergence of interpolated Markov chains to diffusions. Let $\mathbf{b}(\cdot)$ and $\mathbf{b}_\epsilon(\cdot)$, $\epsilon > 0$, be $\mathbb{R}^d$-valued Borel functions on $\mathbb{R}^d$; and let $\boldsymbol{\sigma}(\cdot)$ and $\boldsymbol{\sigma}_\epsilon(\cdot)$, $\epsilon > 0$, be $\mathbb{R}^{d\times d}$ matrix-valued Borel functions on $\mathbb{R}^d$. For each $\epsilon > 0$, let $\{X_k^\epsilon\}$ be an $\mathbb{R}^d$-valued Markov chain with initial condition $X_0^\epsilon = X_0$ such that

$$\mathbf{b}_\epsilon(\mathbf{x}) = (1/\epsilon) E\{X_{k+1}^\epsilon - X_k^\epsilon | X_k^\epsilon = \mathbf{x}\},$$

$$\mathbf{a}_\epsilon(x) = \boldsymbol{\sigma}_\epsilon(\mathbf{x})\boldsymbol{\sigma}_\epsilon'(\mathbf{x}) = (1/\epsilon)\,\mathrm{Cov}\{X_{k+1}^\epsilon - X_k^\epsilon | X_k^\epsilon = \mathbf{x}\}.$$

Interpolate $\{X_k^\epsilon\}$ into a continuous-time process $\{\mathbf{x}^\epsilon(t), t \geq 0\}$ be setting $\mathbf{x}^\epsilon(t) = X_{\lfloor t/\epsilon \rfloor}^\epsilon$, $t \geq 0$. Consider the following conditions:

(K1) $\mathbf{b}(\cdot)$, $\boldsymbol{\sigma}(\cdot)$ are bounded and continuous;

(K2) $\mathbf{b}_\epsilon(\cdot)$, $\boldsymbol{\sigma}_\epsilon(\cdot)$ are uniformly bounded for small $\epsilon > 0$;

(K3) $E\{\sum_{k=0}^{\lfloor t/\epsilon \rfloor} [|\mathbf{b}_\epsilon(X_k^\epsilon) - \mathbf{b}(X_k^\epsilon)|^2 + |\boldsymbol{\sigma}_\epsilon(X_k^\epsilon) - \boldsymbol{\sigma}(X_k^\epsilon)|^2] \cdot \epsilon\} \to 0$, as $\epsilon \to 0$, for all $t > 0$;

(K4) $E\{\sum_{k=0}^{\lfloor t/\epsilon \rfloor} |X_{k+1}^\epsilon - X_k^\epsilon - \mathbf{b}_\epsilon(X_k^\epsilon)\epsilon|^{2+\alpha}\} \to 0$, as $\epsilon \to 0$, for all $t > 0$, for some $\alpha > 0$;

(K5) Let $\mathbf{x}_i(\cdot)$, $i = 1, 2$, be $\mathbb{R}^d$-valued processes, nonanticipative with respect to standard d-dimensional Wiener processes $\mathbf{w}_i(\cdot)$, $i = 1, 2$, respectively. If $(\mathbf{x}_i(\cdot), \mathbf{w}_i(\cdot))$, $i = 1, 2$, satisfy

$$d\mathbf{x}(t) = \mathbf{b}(\mathbf{x}(t))\,dt + \boldsymbol{\sigma}(\mathbf{x}(t))\,d\mathbf{w}(t), \qquad t \geq 0, \tag{11}$$

with $\mathbf{x}(0) = X_0$ in distribution, then the multivariate distributions of $\mathbf{x}_1(\cdot)$ are the same as those of $\mathbf{x}_2(\cdot)$. In other words, (11) has a weakly unique solution; see Ref. 15.

Theorem 3.1. See Ref. 9. Assume (K1)–(K5). Then, $\mathbf{x}^\epsilon(\cdot) \to \mathbf{x}(\cdot)$ weakly in $D^d[0, \infty)$ as $\epsilon \to 0$, where $\mathbf{x}(\cdot)$ satisfies (11).

Now, consider the Metropolis and heat-bath Markov chains with d-dimensional Euclidean state space described in Section 2, and consider the notation introduced therein. To apply Theorem 3.1 to the proof of Theorem 2.1 will require several lemmas.

Let

$$\hat{s}_M(\mathbf{x},\mathbf{y}) = 1, \qquad \text{if } (U_x(\mathbf{x}), \mathbf{y}-\mathbf{x}) \le 0,$$

$$\hat{s}_M(\mathbf{x},\mathbf{y}) = \exp[-(U_x(\mathbf{x}), \mathbf{y}-\mathbf{x})/k_B T], \qquad \text{if } (U_x(\mathbf{x}), \mathbf{y}-\mathbf{x}) > 0,$$

$$\hat{s}_H(\mathbf{x},\mathbf{y}) = 1/2 + (1/4)[1 - \exp[(U_x(\mathbf{x}), \mathbf{y}-\mathbf{x})/k_B T]], \qquad \text{if } (U_x(\mathbf{x}), \mathbf{y}-\mathbf{x}) \le 0,$$

$$\hat{s}_H(\mathbf{x},\mathbf{y}) = [1/2 + (1/4)[1 - \exp[-(U_x(\mathbf{x}), \mathbf{y}-\mathbf{x})/k_B T]]] \times \exp[-(U_x(\mathbf{x}), \mathbf{y}-\mathbf{x})/k_B T], \qquad \text{if } (U_x(\mathbf{x}), \mathbf{y}-\mathbf{x}) > 0,$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Recall how we defined $s(i, \mathbf{x}, y_i)$ in terms of $s(\mathbf{x}, \mathbf{y})$; see Eq. (7). Define $s_M(i, \mathbf{x}, y_i)$, $\hat{s}_M(i, \mathbf{x}, y_i)$, $s_H(i, \mathbf{x}, y_i)$, and $\hat{s}_H(i, \mathbf{x}, y_i)$ analogously in terms of $s_M(\mathbf{x}, \mathbf{y})$, $\hat{s}_M(\mathbf{x}, \mathbf{y})$, $s_H(\mathbf{x}, \mathbf{y})$, and $\hat{s}_H(\mathbf{x}, \mathbf{y})$, respectively. In the sequel, $c_1, c_2, \ldots$ will refer to constants whose value may change from proof to proof.

Lemma 3.1. Assume (A). Then, there exists a constant K such that

$$|s_M(i, \mathbf{x}, y_i) - \hat{s}_M(i, \mathbf{x}, y_i)| \le K|x_i - y_i|^2, \tag{12}$$

$$|s_H(i, \mathbf{x}, y_i) - \hat{s}_H(i, \mathbf{x}, y_i)| \le K|x_i - y_i|^2, \tag{13}$$

for all $\mathbf{x} \in \mathbb{R}^d$, $y_i \in \mathbb{R}$ and $i = 1, \ldots, d$.

Proof. To simplify notation, replace $U(\cdot)/k_B T$ by $U(\cdot)$.
We prove (12) as follows. Let

$$f(\mathbf{x}, \mathbf{y}) = U(\mathbf{y}) - U(\mathbf{x}) - (U_x(\mathbf{x}), \mathbf{y}-\mathbf{x}), \qquad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

By the mean-value theorem and assumption (A),

$$|f(\mathbf{x}, \mathbf{y})| \le c_1 |\mathbf{y}-\mathbf{x}|^2, \qquad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

By considering the four cases corresponding to the possible signs of $U(\mathbf{y}) - U(\mathbf{x})$ and $(U_x(\mathbf{x}), \mathbf{y}-\mathbf{x})$, it can be shown that

$$|s_M(\mathbf{x}, \mathbf{y}) - \hat{s}_M(\mathbf{x}, \mathbf{y})| \le 1 - \exp[-|f(\mathbf{x}, \mathbf{y})|] \le |f(\mathbf{x}, \mathbf{y})| \le c_1 |\mathbf{y}-\mathbf{x}|^2, \tag{14}$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, and (12) follows immediately.

We prove (13) as follows. Using the fact that

$$(1+\alpha)^{-1}=1/2+(1/4)(1-\alpha)+O((1-\alpha)^2), \qquad \text{as } \alpha\to 1,$$

and assumption (A), we get

$$\begin{aligned} s_{\mathrm{H}}(\mathbf{x},\mathbf{y}) &= \frac{1}{1+\exp[U(\mathbf{y})-U(\mathbf{x})]} \\ &= \frac{\exp[U(\mathbf{x})-U(\mathbf{y})]}{1+\exp[U(\mathbf{x})-U(\mathbf{y})]}=\tilde{s}(\mathbf{x},\mathbf{y})+O(|\mathbf{y}-\mathbf{x}|^2), \qquad \text{as } \mathbf{y}\to\mathbf{x}, \end{aligned}$$

where

$$\begin{aligned} \tilde{s}(\mathbf{x},\mathbf{y}) &= 1/2+(1/4)[1-\exp[U(\mathbf{y})-U(\mathbf{x})]], && \text{if } U(\mathbf{y})\le U(\mathbf{x}), \\ \tilde{s}(\mathbf{x},\mathbf{y}) &= [1/2+(1/4)[1-\exp[U(\mathbf{x})-U(\mathbf{y})]]] \\ &\quad \times \exp[U(\mathbf{x})-U(\mathbf{y})], && \text{if } U(\mathbf{y})> U(\mathbf{x}). \end{aligned}$$

Since $s_{\mathrm{H}}(\cdot,\cdot)$ and $\tilde{s}(\cdot,\cdot)$ are bounded, it follows that

$$|s_{\mathrm{H}}(\mathbf{x},\mathbf{y})-\tilde{s}(\mathbf{x},\mathbf{y})|\le c_2|\mathbf{y}-\mathbf{x}|^2, \qquad \mathbf{x},\mathbf{y}\in\mathbb{R}^d.$$

Similarly to the proof of (14), we can show that

$$|\tilde{s}(\mathbf{x},\mathbf{y})-\hat{s}_{\mathrm{H}}(\mathbf{x},\mathbf{y})|\le c_3|\mathbf{y}-\mathbf{x}|^2, \qquad \mathbf{x},\mathbf{y}\in\mathbb{R}^d.$$

Combining these estimates gives

$$|s_{\mathrm{H}}(\mathbf{x},\mathbf{y})-\hat{s}_{\mathrm{H}}(\mathbf{x},\mathbf{y})|\le c_4|\mathbf{y}-\mathbf{x}|^2, \qquad \mathbf{x},\mathbf{y}\in\mathbb{R}^d,$$

and (13) follows immediately. □

The following two lemmas give the crucial estimates of $\mathbf{b}_\epsilon(\cdot)$ and $\boldsymbol{\sigma}_\epsilon(\cdot)$. We shall denote by $N(m,a)(\cdot)$ the scalar normal measure with mean m and variance a. We shall frequently use the trivial estimate

$$\int |\alpha|^n\, dN(0,\epsilon)(\alpha)=O(\epsilon^{n/2}), \qquad \text{as } \epsilon\to 0.$$

Lemma 3.2. Assume (A). Then, the following results hold:

(a) for the Metropolis method,

$$\mathbf{b}_\epsilon(\mathbf{x})=-[U_{\mathbf{x}}(\mathbf{x})/2k_{\mathrm{B}}T]+O(\epsilon^{1/2}), \qquad \text{as } \epsilon\to 0;$$

(b) for the heat-bath method

$$\mathbf{b}_\epsilon(\mathbf{x}) = -[U_\mathbf{x}(\mathbf{x})/4k_\mathrm{B}T] + O(\epsilon^{1/2}), \qquad \text{as } \epsilon \to 0.$$

In both cases, the convergence is uniform for all $\mathbf{x} \in \mathbb{R}^d$.

Proof. To simplify notation replace $U(\cdot)/k_\mathrm{B}T$ by $U(\cdot)$.

The proof of part (a) is as follows. Consider the Metropolis Markov chain. Using Lemma 3.1, we have

$$\begin{aligned}
b_{\epsilon,i}(\mathbf{x}) &= (1/\epsilon) \int (y_i - x_i) p_\epsilon(\mathbf{x}, \mathbf{y})\, d\mathbf{y} \\
&= (1/\epsilon) \int (y_i - x_i) \Bigg(\sum_{k=1}^{d} s_\mathrm{M}(k, \mathbf{x}, y_k) r_\epsilon(x_k, y_k) \prod_{l \neq k} \delta(y_l - x_l) \\
&\qquad\qquad + \gamma_\epsilon(\mathbf{x}) \delta(\mathbf{y} - \mathbf{x}) \Bigg)\, d\mathbf{y} \\
&= (1/\epsilon) \int (y_i - x_i) s_\mathrm{M}(i, \mathbf{x}, y_i)\, dN(x_i, \epsilon)(y_i) \\
&= (1/\epsilon) \int (y_i - x_i) \hat{s}_\mathrm{M}(i, \mathbf{x}, y_i)\, dN(x_i, \epsilon)(y_i) \\
&\quad + (1/\epsilon) \int (y_i - x_i)[s_\mathrm{M}(i, \mathbf{x}, y_i) - \hat{s}_\mathrm{M}(i, \mathbf{x}, y_i)]\, dN(x_i, \epsilon)(y_i) \\
&= (1/\epsilon) \int (y_i - x_i) \hat{s}_\mathrm{M}(i, \mathbf{x}, y_i)\, dN(x_i, \epsilon)(y_i) + O(\epsilon^{1/2}) \\
&= (1/\epsilon^{1/2}) \int_{U_{x_i}(\mathbf{x}) y_i \leq 0} y_i\, dN(0, 1)(y_i) \\
&\quad + (1/\epsilon^{1/2}) \int_{U_{x_i}(\mathbf{x}) y_i > 0} y_i \exp[-U_{\mathbf{x}_i}(\mathbf{x}) y_i \epsilon^{1/2}]\, dN(0, 1)(y_i) \\
&\quad + O(\epsilon^{1/2}), \qquad \text{as } \epsilon \to 0,
\end{aligned} \tag{15}$$

uniformly for $\mathbf{x} \in \mathbb{R}^d$. Obviously,

$$b_{\epsilon,i}(\mathbf{x}) = -U_{\mathbf{x}_i}(\mathbf{x})/2 + O(\epsilon^{1/2}), \qquad \text{as } \epsilon \to 0, \tag{16}$$

uniformly on $\{\mathbf{x}\colon U_{\mathbf{x}_i}(\mathbf{x}) = 0\}$. Assume that $U_{\mathbf{x}_i}(\mathbf{x}) > 0$. Then, completing the square in the second integral in (15) and also using the fact that $U_{\mathbf{x}_i}(\mathbf{x})$ is bounded gives

$$\begin{aligned}
b_{\epsilon,i}(\mathbf{x}) &= (1/\epsilon^{1/2}) \int_{y_i \le 0} y_i \, dN(0,1)(y_i) \\
&\quad + (1/\epsilon^{1/2}) \int_{y_i > 0} y_i \exp[U_{x_i}^2(\mathbf{x})\epsilon/2] \, dN(-U_{x_i}(\mathbf{x})\epsilon^{1/2}, 1)(y_i) + O(\epsilon^{1/2}) \\
&= (1/\epsilon^{1/2}) \int_{y_i \le 0} y_i \, dN(0,1)(y_i) + (1/\epsilon^{1/2}) \int_{y_i > U_{x_i}(\mathbf{x})\epsilon^{1/2}} y_i \, dN(0,1)(y_i) \\
&\quad - U_{x_i}(\mathbf{x}) N(0,1)\{y_i : y_i > U_{x_i}(\mathbf{x})\epsilon^{1/2}\} + O(\epsilon^{1/2}) \\
&= -(1/\epsilon^{1/2}) \int_0^{O(\epsilon^{1/2})} y_i (1/\sqrt{2\pi}) \exp(-y_i^2/2) \, dy_i \\
&\quad - U_{x_i}(\mathbf{x}) \left(1/2 - \int_0^{O(\epsilon^{1/2})} (1/\sqrt{2\pi}) \exp(-y_i^2/2) \, dy_i \right) + O(\epsilon^{1/2}) \\
&= -U_{x_i}(\mathbf{x})/2 + O(\epsilon^{1/2}), \qquad \text{as } \epsilon \to 0,
\end{aligned}$$

uniformly on $\{\mathbf{x}\colon U_{x_i}(\mathbf{x}) > 0\}$, and similarly on $\{\mathbf{x}\colon U_{x_i}(\mathbf{x}) < 0\}$, and hence by (16) for all $\mathbf{x} \in \mathbb{R}^d$. Hence,

$$\mathbf{b}_\epsilon(\mathbf{x}) = -U_{\mathbf{x}}(\mathbf{x})/2 + O(\epsilon^{1/2}), \qquad \text{as } \epsilon \to 0,$$

uniformly for $\mathbf{x} \in \mathbb{R}^d$, as required.

The proof of part (b) involves somewhat more details than part (a), but the method is similar [use (13) instead of (12)]. □

Lemma 3.3. Assume (A). Then, the following results hold:

(a) for the Metropolis method,

$$\boldsymbol{\sigma}_\epsilon(\mathbf{x}) = \mathbf{I} + O(\epsilon^{1/2}), \qquad \text{as } \epsilon \to 0;$$

(b) for the heat-bath method,

$$\boldsymbol{\sigma}_\epsilon(\mathbf{x}) = (1/\sqrt{2})\mathbf{I} + O(\epsilon^{1/2}), \qquad \text{as } \epsilon \to 0.$$

In both cases, the convergence is uniform for all $\mathbf{x} \in \mathbb{R}^d$.

Proof. To simplify notation, replace $U(\cdot)/k_B T$ by $U(\cdot)$.

The proof of part (a) is as follows. Consider the Metropolis Markov chain. Using Lemma 3.2(a) and Lemma 3.1, we have

$$\begin{aligned}
a_{\epsilon,i,j}(\mathbf{x}) &= (1/\epsilon)\int (y_i - x_i - b_{\epsilon,i}(\mathbf{x})\epsilon)(y_j - x_j - b_{\epsilon,j}(\mathbf{x})\epsilon)p_\epsilon(\mathbf{x},\mathbf{y})\,d\mathbf{y}\\
&= (1/\epsilon)\int (y_i - x_i)(y_j - x_j)p_\epsilon(\mathbf{x},\mathbf{y})\,d\mathbf{y} - b_{\epsilon,i}(\mathbf{x})b_{\epsilon,j}(\mathbf{x})\epsilon\\
&= (1/\epsilon)\int (y_i - x_i)(y_j - x_j)\Bigg(\sum_{k=1}^{d} s_{\mathrm{M}}(k,\mathbf{x},y_k)r_\epsilon(x_k,y_k)\prod_{l\neq k}\delta(y_l - x_l)\\
&\quad + \gamma_\epsilon(\mathbf{x})\delta(\mathbf{y}-\mathbf{x})\Bigg)\,d\mathbf{y} + O(\epsilon)\\
&= (1/\epsilon)\int (y_i - x_i)^2 s_{\mathrm{M}}(i,\mathbf{x},y_i)\,dN(x_i,\epsilon)(y_i)\cdot\delta_{ij} + O(\epsilon)\\
&= (1/\epsilon)\int (y_i - x_i)^2 \hat{s}_{\mathrm{M}}(i,\mathbf{x},y_i)\,dN(x_i,\epsilon)(y_i)\cdot\delta_{ij}\\
&\quad + (1/\epsilon)\int (y_i - x_i)^2[s_{\mathrm{M}}(i,\mathbf{x},y_i) - \hat{s}_{\mathrm{M}}(i,\mathbf{x},y_i)]\,dN(x_i,\epsilon)(y_i)\cdot\delta_{ij}\\
&\quad + O(\epsilon)\\
&= (1/\epsilon)\int (y_i - x_i)^2 \hat{s}_{\mathrm{M}}(i,\mathbf{x},y_i)\,dN(x_i,\epsilon)(y_i)\cdot\delta_{ij} + O(\epsilon)\\
&= \int_{U_{x_i}(\mathbf{x})y_i\leq 0} y_i^2\,dN(0,1)(y_i)\cdot\delta_{ij}\\
&\quad + \int_{U_{x_i}(\mathbf{x})y_i>0} y_i^2\exp[-U_{x_i}(\mathbf{x})y_i\epsilon^{1/2}]\,dN(0,1)(y_i)\cdot\delta_{ij}\\
&\quad + O(\epsilon), \qquad \text{as } \epsilon\to 0,
\end{aligned}\tag{17}$$

uniformly for $\mathbf{x}\in\mathbb{R}^d$. Obviously,

$$a_{\epsilon,i,i}(\mathbf{x}) = 1 + O(\epsilon), \qquad \text{as } \epsilon\to 0,\tag{18}$$

uniformly on $\{\mathbf{x}\colon U_{x_i}(\mathbf{x}) = 0\}$. Assume that $U_{x_i}(\mathbf{x}) > 0$. Then, completing the square in the second integral in (17) and also using the fact that $U_{x_i}(\mathbf{x})$ is bounded gives

$$\begin{aligned} a_{\epsilon,i,i}(\mathbf{x}) &= \int_{y_i \leq 0} y_i^2 \, dN(0,1)(y_i) \\ &\quad + \int_{y_i > 0} y_i^2 \exp[U_{x_i}^2(\mathbf{x})\epsilon/2] \, dN(-U_{x_i}(\mathbf{x})\epsilon^{1/2}, 1)(y_i) + O(\epsilon) \\ &= \int_{y_i \leq 0} y_i^2 \, dN(0,1)(y_i) + \int_{y_i > U_{x_i}(\mathbf{x})\epsilon^{1/2}} y_i^2 \, dN(0,1)(y_i) + O(\epsilon^{1/2}) \\ &= 1 - \int_0^{O(\epsilon^{1/2})} y_i^2 (1/\sqrt{2\pi}) \exp(-y_i^2/2) \, dy_i + O(\epsilon^{1/2}) \\ &= 1 + O(\epsilon^{1/2}), \qquad \text{as } \epsilon \to 0, \end{aligned}$$

uniformly on $\{\mathbf{x}\colon U_{x_i}(\mathbf{x}) > 0\}$, and similarly on $\{\mathbf{x}\colon U_{x_i}(\mathbf{x}) < 0\}$, and hence by (18) for all $\mathbf{x} \in \mathbb{R}^d$. Hence,

$$\mathbf{a}_\epsilon(\mathbf{x}) = \mathbf{I} + O(\epsilon^{1/2}), \qquad \text{as } \epsilon \to 0, \tag{19}$$

uniformly for all $\mathbf{x} \in \mathbb{R}^d$.

Now, let $\lambda_{\epsilon,i}(\mathbf{x})$, $i = 1, \dots, d$, be the eigenvalues of $\mathbf{a}_\epsilon(\mathbf{x})$. From (19), we have

$$\det(\lambda I - \mathbf{a}_\epsilon(\mathbf{x})) = (\lambda - 1)^d + (\lambda - 1)^{d-1} O(\epsilon^{1/2}) + \cdots + O(\epsilon^{d/2}),$$

and so

$$|\lambda_{\epsilon,i}(\mathbf{x}) - 1|^d = O(\max\{|\lambda_{\epsilon,i}(\mathbf{x}) - 1|^{d-1}\epsilon^{1/2}, \epsilon^{d/2}\}),$$

and so

$$\lambda_{\epsilon,i}(\mathbf{x}) = 1 + O(\epsilon^{1/2}),$$

and consequently

$$\lambda_{\epsilon,i}^{1/2}(\mathbf{x}) = 1 + O(\epsilon^{1/2}), \qquad \text{as } \epsilon \to 0,$$

uniformly for $\mathbf{x} \in \mathbb{R}^d$. It follows from this that we can choose

$$\boldsymbol{\sigma}_\epsilon(\mathbf{x}) = \mathbf{I} + O(\epsilon^{1/2}), \qquad \text{as } \epsilon \to 0,$$

uniformly for all $\mathbf{x} \in \mathbb{R}^d$, as required.

The proof of part (b) involves somewhat more details than part (a), but the method is similar [use (13) instead of (12)]. □

Proof of Theorem 2.1. To prove part (a), we apply Theorem 3.1 with

$$\mathbf{b}(\cdot) = -U_{\mathbf{x}}(\cdot)/2k_B T \quad \text{and} \quad \boldsymbol{\sigma}(\cdot) = \mathbf{I}.$$

In view of assumption (A) and Lemmas 3.2 and 3.3, conditions (K1) and (K2) are satisfied; furthermore, for every $t>0$,

$$E\left\{\sum_{k=0}^{\lfloor t/\epsilon\rfloor}[|\mathbf{b}_\epsilon(\mathbf{X}_k^\epsilon)-\mathbf{b}(\mathbf{X}_k^\epsilon)|^2+|\boldsymbol{\sigma}_\epsilon(X_k^\epsilon)-\boldsymbol{\sigma}(X_k^\epsilon)|^2]\epsilon\right\}$$

$$=\sum_{k=0}^{\lfloor t/\epsilon\rfloor}O(\epsilon^2)=O(\epsilon),\qquad \text{as } \epsilon\to 0,$$

and so (K3) is satisfied. Now, for $n\geq 0$, we have

$$E\{|X_{k+1}^\epsilon-X_k^\epsilon|^n|X_k^\epsilon=\mathbf{x}\}$$

$$=\int|\mathbf{y}-\mathbf{x}|^n p_\epsilon(\mathbf{x},\mathbf{y})\,d\mathbf{y}=\sum_{i=1}^{d}\int|y_i-x_i|^n s_{\mathrm{M}}(i,\mathbf{x},y_i)\,dN(x_i,\epsilon)(y_i)$$

$$\leq d\int|y_i|^n\,dN(0,\epsilon)(y_i)=O(\epsilon^{n/2}),\qquad \text{as } \epsilon\to 0,$$

uniformly for $\mathbf{x}\in\mathbb{R}^d$. Hence, using the uniform boundedness of $\mathbf{b}_\epsilon(\cdot)$, for every $t>0$,

$$E\left\{\sum_{k=0}^{\lfloor t/\epsilon\rfloor}|X_{k+1}^\epsilon-X_k^\epsilon-\mathbf{b}_\epsilon(X_k^\epsilon)\epsilon|^4\right\}$$

$$=\sum_{k=0}^{\lfloor t/\epsilon\rfloor}O(\epsilon^2)=O(\epsilon),\qquad \text{as } \epsilon\to 0,$$

and so (K4) is satisfied. Finally, it is well known that (K5) is satisfied under assumption (A); see Ref. 15. Part (a) [and similarly part (b)] now follows from Theorem 3.1. □

References

1. CERNY, V., *A Thermodynamical Approach to the Travelling Salesman Problem*, Journal of Optimization Theory and Applications, Vol. 45, pp. 41–51, 1985.
2. ALUFFI-PENTINI, F., PARISI, V., and ZIRILLI, F., *Global Optimization and Stochastic Differential Equations*, Journal of Optimization Theory and Applications, Vol. 47, pp. 1–16, 1985.
3. GIDAS, B., *Global Optimization via the Langevin Equation*, Proceedings of the Twenty-Fourth IEEE Conference on Decision and Control, Fort Lauderdale, Florida, pp. 774–778, 1985.
4. COLLINS, N. E., EGLESE, R. W., and GOLDEN, B. L., *Simulated Annealing—An Annotated Bibliography*, American Journal of Mathematical and Management Sciences, Vol. 8, pp. 209–307, 1988.

5. BROOKS, D. G., and VERDINI, W. A., *Computational Experience with Generalized Simulated Annealing over Continuous Variables*, American Journal of Mathematical and Management Sciences, Vol. 8, pp. 425–449, 1988.
6. BINDER, K., *Monte Carlo Methods in Statistical Physics*, Springer-Verlag, Berlin, Germany, 1978.
7. KUSHNER, H. J., and CLARK, D., *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, Berlin, Germany, 1978.
8. KUSHNER, H. J., *Approximation and Weak Convergence Methods for Random Processes*, MIT Press, Cambridge, Massachusetts, 1984.
9. KUSHNER, H. J., *On the Weak Convergence of Interpolated Markov Chains to a Diffusion*, Annals of Probability, Vol. 2, pp. 40–50, 1974.
10. CHUNG, K. L., *Markov Processes with Stationary Transition Probabilities*, Springer-Verlag, Heidelberg, Germany, 1960.
11. BILLINGSLEY, P., *Convergence of Probability Measures*, Wiley, New York, New York, 1968.
12. HAJEK, B., *Cooling Schedules for Optimal Annealing*, Mathematics of Operations Research, Vol. 13, pp. 311–329, 1988.
13. CHIANG, T. S., HWANG, C. R., and SHEU, S. J., *Diffusion for Global Optimization in* $\mathbb{R}^n$, SIAM Journal on Control and Optimization, Vol. 25, pp. 737–752, 1987.
14. GELFAND, S. B., and MITTER, S. K., *Simulated Annealing-Type Algorithms for Multivariate Optimization*, Algorithmica (in press).
15. GIKHMAN, I. I., and SKOROHOD, A. V., *Stochastic Differential Equations*, Springer-Verlag, Berlin, Germany, 1972.