JOURNAL OF OPTIMIZATION THEORY AND APPLICATIONS: Vol. 100, No. 3, pp. 499 513, MARCH 1999

A Strong Approximation Theorem for Stochastic Recursive Algorithms[1,2]

V. S. BORKAR[3] AND S. K. MITTER[4]

Communicated by W. B. Gong and D. D. Yao

Abstract. The constant stepsize analog of Gelfand–Mitter type discrete-time stochastic recursive algorithms is shown to track an associated stochastic differential equation in the strong sense, i.e., with respect to an appropriate divergence measure.

Key Words. Stochastic algorithms, approximation of stochastic differential equations, constant stepsize algorithms, asymptotic behavior.

1. Introduction

In Refs. 1 and 2, Gelfand and Mitter studied the following stochastic recursive algorithm for finding a global minimum of a smooth function $U: R^d \to R$:

$$X_{k+1} = X_k - a_k(\nabla U(X_k) + \xi_k) + \sqrt{a_k} b_k W_k, \qquad k \geq 0. \tag{1}$$

Here, $\{\xi_k\}$ is a sequence of random variables representing measurement noise; $\{W_k\}$ is i.i.d. $N(0, I)$ noise (I being the identity matrix) added deliberately to avoid being trapped in local minima. The stepsize sequence $\{a_n\}$ is of the usual stochastic approximation algorithm, i.e., it is square-summable, but not summable. The sequence $\{b_n\}$ is another positive decreasing sequence chosen appropriately. Gelfand and Mitter show that (1) tracks in

[1]This paper is dedicated by the second author to his friend Larry Ho on the occasion of his 65th birthday for his friendship, support, and wisdom over many years.

[2]This work was supported by US Army Research Grant DAAL03-92-G-0115, Center for Intelligent Control Systems, MIT, Cambridge, Massachusetts.

[3]Associate Professor, Department of Computer Science and Automation, Indian Institute of Science, Bangalore, India.

[4]Professor, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts.

0022-3239/99/0300-0499$16.00/0

an appropriate sense the continuous-time simulated annealing algorithm, or Langevin algorithm, given by the stochastic differential equation

$$dX(t) = -\nabla U(X(t))\, dt + \epsilon(t) dW(t), \tag{2}$$

where $W(\cdot)$ is a standard Brownian motion in R^d and $\epsilon(t)$ is a positive function decreasing to zero at a suitable rate. Then, (1) asymptotically mimics the asymptotic behavior of (2), which is to converge in probability to the set of global minima of $U(\cdot)$.

There are situations, however, when one would like to replace $\{a_n\}$, $\{b_n\}$ by constants $a, b > 0$. This is often done for ease of hardware implementations when the algorithms are hard-wired. Even more significantly, one does so when the environment [i.e., $U(\cdot)$] is actually varying slowly in time and one is expected to track its global minimum. However, with constant a, b the algorithm (1) will not converge in general to a point, even in probability. For example, if $\{\xi_n\}$ are i.i.d., (1) is a time-homogeneous Markov process which will converge to a stationary distribution at best. Therefore, one has to rephrase the desired behavior of the constant stepsize algorithm as follows: its limiting distributions should be concentrated near the global minima.

This is usually achieved in two steps. The first step is to show that, as $a \to 0$, the discrete recursion tracks an associated stochastic differential equation better and better. In particular, its limiting distributions remain close to the invariant probability distribution of the Markov process generated by the stochastic differential equation. The second step is to show that the latter, in turn, have the desired limiting behavior (e.g., concentrate on global minima) in the small-noise limit $b \to 0$. Traditionally, both these limits are in the weak sense, i.e., in the sense of weak convergence of probability measures (Ref. 3, Chapter 2). The aim of this paper is to show a stronger approximation theorem for the first step, viz., in the sense of information theoretic (Kulback–Leibler) divergence. This is useful, because usually it is only the invariant distributions of the s.d.e. that are analytically accessible. As for the second step, i.e., the small-noise limit, such a strong approximation is usually not possible because the limit of invariant distributions as $b \to 0$ is usually singular with respect to any of them.

We should also add that our proof technique provides a stochastic analog of the Hirsch lemma (Ref. 4, Theorem 1) for ordinary differential equations, which gives conditions under which perturbed trajectories of the same track the original asymptotic behavior closely. This proof technique is of independent interest. In addition to yielding a stronger approximation result, it has the added advantage of not requiring the asymptotic stationarity of the discrete recursions or an a priori condition of tightness on their limiting distributions as the stepsize decreases to zero. In addition, the way

our proof is structured, it is likely to provide a useful starting point for a quantitative estimation of the approximation error involved.

We shall consider a somewhat more general situation than (1) and (2), not restricting necessarily to a gradient search. Thus, we consider a more general discrete recursion given by

$$X_{k+1} = X_k + a(h(X_k) + M_k) + \sqrt{a}bW_k, \tag{3}$$

where $h: R^d \to R$ is Lipschitz, X_0 is prescribed in law with $E[\|X_0\|^m] < \infty$, $\forall m$, $\{W_k\}$ is as before, and $\{M_k\}$ is a martingale difference sequence representing the measurement noise. We compare (3) with the stochastic differential equation

$$dX(t) = h(X(t))\,dt + b\,dW(t), \tag{4}$$

with $W(\cdot)$ as before. Our assumptions, spelled out in the next section, ensure that (4) has a unique invariant probability measure $\eta_b(\cdot) \in P(R^d)$; here and later on, $P(\ldots)$ stands for the Polish space of probability measures on the Polish space "..." with Prohorov topology (see Ref. 3, Chapter 2). The aim is to show that the laws of X_k as $k \to \infty$ converge to a neighborhood of η_b, in the strong sense alluded to earlier, for sufficiently small a. The next section spells out the assumptions under which this will be achieved. Section 3 proves an approximation result for approximating (4) by a suitably interpolated version of (3) on finite time intervals. Section 4 uses this result to prove the main convergence theorem.

2. Preliminaries

We shall state our assumptions, including for convenience the ones that were already mentioned in passing in the Introduction. Before doing so, we recast (3) as an Itô differential equation. Since $\{W_k\}$ are i.i.d. $N(0, I)$, we may view them as Brownian increments. Specifically, we postulate a d-dimensional standard Brownian motion $\tilde{W}(\cdot)$ such that

$$\tilde{W}((k+1)a) - \tilde{W}(ka) = \sqrt{a}W_k, \qquad k \geq 0. \tag{5}$$

This may require an enlargement of the underlying probability space, but that does not affect our analysis. Our first assumption is:

(A1) $h(\cdot): R^d \to R^d$ is Lipschitz and bounded.

This ensures a unique strong solution to (4); see Ref. 5.

Define

$$[t]_a = ka, \qquad \text{where } ka \leq t < (k+1)a,$$

for $t \geq 0$, $a > 0$, and

$$\xi_t = M_k, \quad \text{for } ka \leq t < (k+1)a, \qquad k \geq 0.$$

Then, the process $\tilde{X}(\cdot)$ defined by

$$\tilde{X}(t) = X_0 + \int_0^t (h(\tilde{X}([s]_a)) + \xi_s)\, ds + b\tilde{W}(t) \tag{6}$$

for $t \geq 0$ satisfies, by construction,

$$\tilde{X}(ka) = X_k, \qquad k \geq 0.$$

Our next assumption concerns $\{M_n\}$. Let

$$\mathcal{F}_n = \sigma(X_k, k \leq n; M_j, W_j, j < n), \qquad n \geq 0.$$

Then:

(A2) $\{M_n\}$ is a sequence of R^d-valued integrable random variables satisfying

- (i) $\sup_n E[\|M(n)\|^2] \triangleq C < \infty$,
- (ii) $E[M_n / \mathcal{F}_n] = 0$,
- (iii) W_n is independent of $\mathcal{F}_n \vee \sigma(M_n)$, $\forall n$,
- (iv) $E[\|M_n\| / \mathcal{F}_n] \leq C_1(1 + \|X_n\|)$, for some constant $C_1 > 0$.

Condition (ii) characterizes $\{M_n, \mathcal{F}_n\}$ as a martingale difference sequence. Combining (A1) with (iv) above, a standard argument using the Gronwall inequality shows that $\tilde{X}(\cdot)$, and therefore $\{M_n\}$, has bounded moments at all times, bounded uniformly on every compact time interval. The same obviously holds true for $X(\cdot)$. The next assumption is:

(A3) $X(\cdot)$ is stable, i.e., positive recurrent (Ref. 6).

Assumption (A1) is sufficient to ensure that (4) will have strictly positive transition probability densities for positive times. Assumption (A3) then ensures that it will be ergodic with a unique invariant probability measure $\eta_b(\cdot)$, which has a strictly positive density (Ref. 6).

Sufficient conditions for Assumption (A3) to hold can be given in terms of stochastic Liapunov functions. As an example, we cite one from Ref. 7:

(C1) There exists a twice continuously differentiable function $w: R^d \to R^+$ such that

- (i) $\lim_{\|x\| \to \infty} w(x) = \infty$ uniformly in $\|x\|$,
- (ii) $w(x)$, $\|\nabla w(x)\|$ have polynomial growth,

(iii) there exists $a_1 > 0$, $\epsilon > 0$ such that, for $\|x\| > a_1$,

$$\sum_i h_i(x) \frac{\partial w}{\partial x_i}(x) + (1/2)b^2 \sum_i \frac{\partial^2 w}{\partial x_i^2}(x) \leq -\epsilon,$$

$$\|\nabla w(x)\|^2 \geq b^2.$$

See Ref. 7, Section 5, for details. Our next assumption is:

(A4) There exists a $c > 0$ such that, for all $b > 0$,

$$\int \eta_b(dx)\|x\|^{2+c} < \infty,$$

and $\eta_0 = \lim_{b \to 0} \eta_b$ is well defined, the limit being in $P(R^d)$.

We also need a stability condition on (3):

(A5) For sufficient small a, (3) is stable, i.e., the laws of X_k, $k > 0$, remain tight. As in traditional stochastic approximation theory, sufficient conditions for such stability to hold tend to be problem specific. Important examples are the Liapunov and perturbed Liapunov methods (Ref. 8). Since (3) is an algorithm, we can modify the dynamics if need be and, in particular, ensure a Liapunov-type condition by modifying $h(x)$ for very large $\|x\|$ by incorporating a suitable penalty term.

By way of illustration, we state below one Liapunov-type condition for stability of (3).

(C2) This condition is the same as (C1), except that (iii) is replaced by:

(iii′) There exist R, a_1, $\epsilon > 0$, such that, for $\|x\| > a_1$,

$$\sum_i h_i(\bar{x}) \frac{\partial w}{\partial x_i}(x) + (1/2)b^2 \sum_i \frac{\partial^2 w}{\partial x_i^2}(x) \leq -\epsilon,$$

for all $\bar{x}$ satisfying $\|x - \bar{x}\| \leq R\|x\|$.

Lemma 2.1. Under Condition (C2), Assumption (A5) holds.

Sketch of Proof. Fix $k \geq 1$. Let

$$A_1 = \left\{ \inf_{t \in [ka,(k+1)a]} \|\tilde{X}(t)\| \leq a_1 \right\},$$

$$A_2 = \left\{ \sup_{t \in [ka,(k+1)a]} \|\tilde{X}(t) - \tilde{X}(ka)\| > R\|\tilde{X}(ka)\| \right\}.$$

Then,

$$E[(w(\tilde{X}((k+1)a)-w(\tilde{X}(ka)))I\{\|\tilde{X}(ka)\|>2a_1\}/F_{ka})]$$
$$=E\left[\int_{ka}^{(k+1)a}\left(\langle\nabla w(\tilde{X}(s)),h(\tilde{X}([s]_a))\rangle\right.\right.$$
$$\left.\left.+(1/2)b^2\sum_i\frac{\partial^2 w}{\partial x_i^2}(\tilde{X}(s))\right)ds/F_{ka}\right]I\{\|\tilde{X}(ka)\|>2a_1\}$$
$$\leq -\epsilon a+KP(A_1\cup A_2/\|\tilde{X}(ka)\|>2a_1),$$

for a suitable constant $K>0$. It is not difficult to show that the second term on the r.h.s. can be made less than $\epsilon a/2$ by making a sufficiently small. The rest of the proof then follows as in Ref. 9. □

Letting $\mathcal{L}(\ldots)$ stand for "the law of ...," Assumption (A5) allows us to postulate a compact set $\mathcal{C}\subset P(R^d)$ such that $\mathcal{L}(\tilde{X}(t))\in\mathcal{C}$, $\forall t\geq 0$.

We conclude this section with the observation that, in many cases, condition (i) of Assumption (A2) follows from condition (iv) of Assumption (A2) and a Liapunov-type condition imposed to ensure Assumption (A5).

3. Approximation on Finite Time Intervals

The aim of this section is to show that $\mathcal{L}(\tilde{X}(t))$ approximates $\mathcal{L}(X(t))$ in a strong sense, uniformly on compact time intervals. Set $X(0)=X_0$ and $W(\cdot)=\tilde{W}(\cdot)$. That is, we solve (4) with these specifications. This is possible because of the existence and uniqueness of a strong solution to (4). Let L be the Lipschitz constant for $h(\cdot)$, let C be the constant in Assumption (A2) (i), and let

$$K_T=\sup_{t\in[0,T]}E[\|\tilde{X}(t)\|^2],\qquad \text{for } T>0.$$

Lemma 3.1. $\forall t$, $E[\|\tilde{X}(t)-X(t)\|^2]\to 0$ as $a\to 0$.

Proof. Fix $T>0$. Let $N=$ the least integer exceeding T/a. Then, for $t\in[0,T]$, (4) and (6) lead to

$$E[\|X(t)-\tilde{X}(t)\|^2]$$
$$\leq L\left(E\left[\int_0^t\|X(s)-\tilde{X}(s)\|^2\,ds\right]+E\left[\int_0^t\|\tilde{X}(s)-\tilde{X}([s]_a)\|^2\,ds\right]\right)$$
$$+\sum_{n=0}^{N}a^2(C+b^2d). \tag{7}$$

From (6), we have

$$E[\|\tilde{X}(s)-\tilde{X}([s]_a)\|^2]\leq a^2(LK_T+C+b^2d).$$

Thus, (7) leads to

$$\begin{aligned}&E[\|X(t)-\tilde{X}(t)\|^2]\\&\leq L\int_0^t E[\|X(s)-\tilde{X}(s)\|^2]\,ds\\&+a^2T(LK_T+C+b^2)L+a(T+1)(C+b^2d),\end{aligned}$$

where we use the fact that $aN\in[T, T+1]$. By the Gronwall inequality, it follows that $E[\|X(t)-\tilde{X}(t)\|^2]$ is $O(a)$ for $t\in[0, T]$. Since $T>0$ was arbitrary, the claim follows. □

Let $p(x, t)$, $\tilde{p}(x, t)$ denote the probability densities of $X(t)$, $\tilde{X}(t)$, respectively, for $t>0$. Let $T>0$, $T_0\geq 1$ and $\mathcal{L}(X_0)=\mu\in\mathcal{C}$.

Lemma 3.2. There exist constants $\alpha(1)$, $\alpha(2)$, $\beta(1)$, $\beta(2)>0$ such that

$$\alpha(1)\exp(-\beta(1)\|x\|^2)\leq p(x, t), \tilde{p}(x, t)\leq\alpha(2)\exp(-\beta(2)\|x\|^2),$$

for all $x\in R^d$, $t\in[T_0, T_0+T]$, and any choice of $\mu\in\mathcal{C}$.

Proof. For $t>s$, x, $y\in R^d$, let $p(y, s; x, t)$ denote the transition probability density of the Markov process $X(\cdot)$. Then, by the estimates of Ref. 10, there exists constants c_1, c_2, c_3, $c_4>0$ such that

$$\begin{aligned}&c_3t^{-d/2}\exp(-c_4\|x-y\|^2/t)\\&\leq p(y, 0; x, t)\\&\leq c_1t^{-d/2}\exp(-c_2\|x-y\|^2/t),\end{aligned}$$

for $0<t\leq T_0+T$. Thus, for $t\in[T_0, T_0+T]$,

$$p(x, t)\leq c_1T_0^{-d/2}\exp(-c_2(T_0+T)^{-1}\|x\|^2)\varphi(x),$$

where

$$\varphi(x)=\int\exp(-c_2(T_0+T)^{-1})(2\langle x, y\rangle-\|y\|^2)\mu(dy).$$

Now, for any $b>0$,

$$\lim_{\|x\|\to\infty}\exp(-b\|x\|^2)\varphi(x)=0.$$

The limit is uniform over $\mu \in \mathscr{C}$ by Dini's theorem. Thus, $\varphi(x)$ is bounded by $K_b \exp(b\|x\|^2)c_2(T_0+T)^{-1}/2$ for any $b>0$ and a corresponding constant $K_b>0$. Pick

$$b = c_2(T_0+T)^{-1}/2,$$

and set

$$\alpha(2) = K_b c_1 T_0^{-d/2}, \qquad \beta(2) = c_2(T_0+T)^{-1}/2,$$

to obtain the upper bound on $p(x, t)$, uniform in $x \in R^d$, $t \in [T_0, T_0+T]$, and $\mu \in \mathscr{C}$. The lower bound is obtained similarly. For $\tilde{p}(x, t)$, use the fact that the one-dimensional marginals $\mathscr{L}(\tilde{X}(t))$, $t \geq 0$, of the non-Markov process $X(\cdot)$ in fact equal the one-dimensional marginals $\mathscr{L}(\hat{X}(t))$, $t \geq 0$, of a Markov process $\hat{X}(\cdot)$ satisfying a stochastic differential equation similar to (4), but with $h(\cdot)$ replaced by a suitable measurable function $\hat{h}(\cdot)$: $R^d \to R^d$ (Refs. 11 and 12). The above arguments can then be applied to $\hat{X}(\cdot)$. □

Corollary 3.1. $\tilde{p}(\cdot, \cdot) \to p(\cdot, \cdot)$ as $a \to 0$ in $C(R^d \times [T_0, T_0+T])$, uniformly with respect to $\mu \in \mathscr{C}$.

Proof. By localizing the estimates of Ref. 13 (Theorem 1.1, p. 419), we conclude that, for some $\bar{\alpha} > 0$, $\tilde{p}(x, t)$ is Hölder continuous with exponent $\bar{\alpha}$ in x and with exponent $\bar{\alpha}/2$ in t. We use once again the fact that $\tilde{p}(x, t)$ is also the probability density of $\hat{X}(t)$, $\hat{X}(\cdot)$ a Markov process (Refs. 11 and 12) and, hence, satisfies the appropriate parabolic p.d.e., viz., the associated forward Kolmogorov equation. Moreover, its Hölder constant in any bounded open subset of R^d depends only on the bound on $\tilde{p}(x, t)$ on that set. By the preceding lemma, this bound is uniform in x, t, μ. Therefore, the family $\{\tilde{p}(\cdot, \cdot); a>0, \mu \in \mathscr{C}\} \subset C(R^d \times [T_0, T_0+T])$ is bounded equicontinuous and, by the Arzela–Ascoli theorem, relatively compact. Let $\phi(\cdot, \cdot)$ be a limit point thereof. Then, in particular, for fixed t and μ, $\tilde{p}(x, t) \to \phi(x, t)$ uniformly on compacts as $a \to 0$ along a certain subsequence. Since $\tilde{p}(\cdot, t)$, t satisfies the bound of the preceding lemma, one can invoke the dominated convergence theorem to conclude that, for $f \in C_b(R^d)$,

$$E[f(\tilde{X}(t))] = \int f(x)\tilde{p}(x, t)\,dt \to \int f(x)p(x, t)\,dt,$$

as $a \to 0$ along the appropriate subsequence. However, by Lemma 3.1, $\tilde{X}(t) \to X(t)$ in law and, therefore,

$$E[f(\tilde{X}(t))] \to E[f(X(t))], \qquad \text{as } a \to 0.$$

Thus,

$$\phi(x, t) = p(x, t), \qquad \forall x, t,$$

completing the proof. □

Let $q_b(\cdot)$ denote the density of $\eta_b(\cdot)$ with respect to the Lebesgue measure on R^d.

Corollary 3.2. $\int q_b(x) \log(q_b(x)/\tilde{p}(x, t))\, dx \to$ $\int q_b(x) \log(q_b(x)/p(x, t))\, dx$, as $a \to 0$, uniformly in $t \in [T_0, T_0 + T]$ and $\mu \in \mathscr{C}$.

Proof. We need to show that

$$\int q_b(x) \log(p(x, t)/\tilde{p}(x, t))\, dx \to 0,$$

uniformly in $t \in [T_0, T_0 + T]$, $\mu \in \mathscr{C}$.

This follows easily by combining the foregoing with the observation that $|\log(p(x, t)/\tilde{p}(x, t))|$ has a uniform quadratic bound in x, uniform in $t \in [T_0, T_0 + T]$ and $\mu \in \mathscr{C}$, using the first half of Assumption (A4). □

4. Main Results

We shall proceed through a sequence of lemmas. For a probability density $\hat{p}(\cdot)$ on R^d, define

$$V(\hat{p}(\cdot)) = \int q_b(x) \log(q_b(x)/\hat{p}(x))\, dx, \tag{8}$$

with $+\infty$ a possible value for the integral.

Lemma 4.1. $V(q_b(\cdot)) = 0$, $V(\hat{p}(\cdot)) \geq 0$ and $= 0$ if, and only if, $\hat{p}(\cdot) = q_b(\cdot)$ a.e. Furthermore, $V(p(\cdot, t))$ is strictly decreasing in t as long as $p(\cdot, t) \neq q_b(\cdot)$.

Proof. Note that (8) defines $V(\hat{p}(\cdot))$ as the information theoretic divergence between $q_b(\cdot)$ and $\hat{p}(\cdot)$. The first sentence of the lemma is immediate. The second also follows exactly as in the discrete case; see for example, Ref. 14 (pp. 34). We include the details for the sake of completeness. Let $t > s$ and let $X^1(\cdot), X^2(\cdot)$ be the solutions of (4) with initial law $\mu(dx) \in P(R^d)$ and $\eta_b(dx)$, respectively, with $\mu \neq \eta_b$. Let $p^1(x, y)$, $p^1(x)$, $\bar{p}^1(y)$, $p^1(y \mid x)$,

$\bar{p}^1(x|y)$ denote, respectively, the joint density of $(X^1(s), X^1(t))$, density of $X^1(s)$, density of $X^1(t)$, conditional density of $X^1(t)$ given $X^1(s)$, and conditional density of $X^1(s)$ given $X^1(t)$, respectively. Let $p^2(x, y)$, $p^2(x)$, $\bar{p}^2(y)$, $p^2(y|x)$, $\bar{p}^2(x|y)$ denote the corresponding entities for $X^2(\cdot)$. Then,

$$p^2(x) = \bar{p}^2(x) = q_b(x) \text{ and } p^1(y|x) = p^2(y|x), \qquad \forall x, y.$$

Therefore,

$$\begin{aligned}
&\iint dx\, dy\, p^2(x, y) \log(p^2(x, y)/p^1(x, y)) \\
&= \iint dx\, dy\, p^2(x) p^2(y|x) \log(p^2(x) p^2(y|x)/p^1(x) p^1(y|x)) \\
&= \iint dx\, dy\, p^2(x) p^2(y|x) \log(p^2(x)/p^1(x)) \\
&= \int dx\, p^2(x) \log(p^2(x)/p^2(x)) = V(p^1(\cdot)).
\end{aligned}$$

Also,

$$\begin{aligned}
&\iint dx\, dy\, p^2(x, y) \log(p^2(x, y)/p^1(x, y)) \\
&= \iint dx\, dy\, \bar{p}^2(y) \bar{p}^2(x|y) \log(\bar{p}^2(y) \bar{p}^2(x|y)/\bar{p}^1(y) \bar{p}^1(x|y)) \\
&= \int dy\, \bar{p}^2(y) \log(\bar{p}^2(y)/\bar{p}^1(y)) \\
&\quad + \iint dx\, dy\, \bar{p}^2(y) \bar{p}^2(x|y) \log(\bar{p}^2(xy)/\bar{p}^1(x|y)).
\end{aligned}$$

The first term of the right-hand side is $V(\bar{p}^1(\cdot))$. The second term is nonnegative and is in fact strictly positive unless

$$\bar{p}^2(y|x) = \bar{p}^1(y|x), \qquad \text{a.e. } (x, y).$$

Since

$$\bar{p}^i(y|x) = p^i(x|y) p^i(y)/\bar{p}^i(x), \qquad i = 1, 2,$$

this reduces to

$$\bar{p}^1(y)/\bar{p}^2(y) = p^1(x)/p^2(x), \qquad \text{for a.e. } (x, y).$$

By continuity, the qualification "a.e." can be dropped from this equality. Clearly, both sides must equal a constant. However,

$$\int \bar{p}^i(y)\, dy = 1, \qquad \text{for } i = 1, 2,$$

implying

$$\bar{p}^1(\cdot) = \bar{p}^2(\cdot) = q_b(\cdot),$$

a contradiction. It follows that

$$V(p^1(\cdot)) = \iint dx\, dy\, p^2(x, y) \log(p^2(x, y)/p^1(x, y)) > V(\bar{p}^1(\cdot)),$$

proving the claim. □

Let T, T_0 be as before, and let Γ = the closure in $P(R^d)$ of $\{\mathcal{L}(X(t)) | t \in [T_0, T_0 + T], \mathcal{L}(X(0)) \in \mathcal{C}\}$. In view of the estimates of Lemma 3.2, it follows that Γ is tight and therefore compact in $P(R^d)$. Let $\epsilon > 0$ and denote by B_ϵ the set of $\mu \in P(R^d)$ having a density $\hat{p}(\cdot)$ that satisfies $V(\hat{p}(\cdot)) < \epsilon$. Let $\Gamma_\epsilon = \Gamma \setminus B_\epsilon$.

Lemma 4.2. There exists a $\Delta > 0$ such that the following holds. Whenever $\mathcal{L}(X(0)) \in \mathcal{C}$ and $\mathcal{L}(X(t)) \in \Gamma_\epsilon \; \forall t \in [T_0, T_0 + T]$, one has

$$V(p(\cdot, T_0 + T)) < V(p(\cdot, T_0)) - \Delta. \tag{9}$$

Proof. Suppose (9) does not hold. Then, there exist solutions $X^n(\cdot)$ of (4) with initial laws $\mu_n \in \mathcal{C}$, $n = 1, 2, \ldots$, such that the corresponding laws $\mathcal{L}(X^n(t))$ have densities $p^n(\cdot, t)$, $t \geq 0$, satisfying

$$V(p^n(\cdot, T_0)) \geq V(p^n(\cdot, T_0 + T)) \geq V(p^n(\cdot, T_0)) - 1/n,$$
$$\mathcal{L}(X^n(t)) \in \Gamma_\epsilon, \qquad t \in [T_0, T_0 + T], \tag{10}$$

for $n = 1, 2, \ldots$. Since $\mathcal{C}$ is compact, we may drop to a subsequence if necessary and suppose that $\mu_n \to \mu_\infty$ in $\mathcal{C}$. Now, the map

$$\mathcal{L}(X(0)) \in P(R^d) \to \mathcal{L}(X(\cdot)) \in P(C((0, \infty), R^d)),$$

defined by (4), is continuous. Thus, along the above subsequence,

$$\mathcal{L}(X^n(\cdot)) \to \mathcal{L}(X^\infty(\cdot)), \qquad \text{in } P(C((0, \infty); R^d)),$$

for a process $X^{\varkappa}(\cdot)$ satisfying (4) with $\mathcal{L}(X^{\infty}(0))=\mu_{\varkappa}$. In particular,

$$\mathcal{L}(X^n(t)) \to \mathcal{L}(X^{\infty}(t)), \qquad \text{for } t\in[T_0, T_0+T].$$

As in Corollary 3.1, $p^n(\cdot, t)\to p^{\infty}(\cdot, t)$ uniformly on compacts, where $p^{\infty}(\cdot, t)$ is the density of $\mathcal{L}(X^{\infty}(t))$. It follows that $\mathcal{L}(X^{\infty}(t))\in\Gamma_{\epsilon}$ for $t\in[T_0, T_0+T]$. Furthermore, passing to the limit in (10), we get

$$V(p^{\infty}(\cdot, T_0))=V(p^{\infty}(\cdot, T_0+T)),$$

a contradiction to Lemma 4.1. The claim follows. □

Our main result is the following theorem.

Theorem 4.1. For any $\delta>0$, $\mathcal{L}(X_n)\to B_{\delta}$ for sufficiently small a.

Proof. It suffices to prove that $\mathcal{L}(\tilde{X}(t))\to B_{\delta}$ for sufficiently small a. Choose a so small that, for a given $\nu\in(0, \Delta/6)$, we have

$$\sup_{t\in[T_0, T_0+T]}\sup_{\mu\in\mathcal{G}}\left|\int q_0(x)\log(p(x, t)/\tilde{p}(x, t))\,dx\right|<\nu. \tag{11}$$

This is possible by Corollary 3.2. For $n\geq 0$, let $\bar{X}^n(t)$, $t\in[nT, T_0+(n+1)T]$, denote the solutions of (4) with initial law $\mathcal{L}(\bar{X}^n(nT))=\mathcal{L}(\tilde{X}(nT))$. Let $\bar{p}^n(\cdot, t)$ denote the probability density of $X^n(t)$ for $t\in[nT, T_0+(n+1)T]$. By Lemma 4.2,

$$V(\bar{p}^n(\cdot, T_0+(n+1)T))\leq V(\bar{p}^n(\cdot, T_0+nT))-\Delta, \tag{12}$$

whenever

$$\mathcal{L}(\bar{X}^n(t))\notin B_{\epsilon}, \qquad \forall t\in[T_0+nT, T_0+(n+1)T]. \tag{13}$$

By (11),

$$V(\tilde{p}(\cdot, T_0+(n+1)T))\leq V(\tilde{p}(\cdot, T_0+nT))-(\Delta-2\nu), \tag{14}$$

whenever (12) holds. If $\mathcal{L}(\tilde{X}(t))$, $t\in[T_0+nT, T_0+(n+1)T]$, does not intersect $B_{\epsilon+\nu}$, it follows from (11) that (12) must hold and therefore (13) and (14) must hold. However, (14) can hold for at most finitely many consecutive values of n. Thus eventually, $\mathcal{L}(\tilde{X}(t))$ and $\mathcal{L}(X^n(t))$, for $t\in[t_0+nT, T_0+(n+1)T]$, must intersect $B_{\epsilon+\nu}$, B_{ϵ}, respectively. Now, for all n,

$$V(\bar{p}^n(\cdot, t))\leq V(\bar{p}^n(\cdot, s)), \qquad \forall t\geq s \text{ in } [T_0+nT, T_0+(n+1)T].$$

Hence, for all n,

$$V(\tilde{p}(\cdot, t))\leq V(\tilde{p}(\cdot, s))+2\nu, \qquad \forall t\geq s \text{ in } [T_0+nT, T_0+(n+1)T].$$

That is, if $\mathcal{L}(\tilde{X}(t))$, $t\in[T_0+nT, t_0+(n+1)T]$, intersects $B_{\epsilon+\nu}$, it remains in $B_{\epsilon+3\nu}$. However, since

$$3\nu<\Delta-2\nu,$$

(14) ensures that $\mathcal{L}(\tilde{X}(t))$, $t\in[T_0+(n+1)T, T_0+(n+2)T]$, will intersect $B_{\epsilon+\nu}$ again. It follows that $\mathcal{L}(\tilde{X}(t))$ remains in $B_{\epsilon+5\nu}$ once it hits $B_{\epsilon+\nu}$, which it does in finite time. Choose, $\epsilon, \nu>0$ such that $\epsilon+5\nu<\delta$. This concludes the proof. □

Suppose that $\eta_b\to\eta$ in $P(R^d)$ as $b\to 0$ and let

$$G_\epsilon=\{\mu\in P(R^d)|\rho(\mu,\eta)<\epsilon\},$$

where ρ is the Prohorov metric on $P(R^d)$ (Ref. 3, Chapter 2) and $\epsilon>0$ is arbitrary.

Corollary 4.1. Given $\epsilon>0$, $\mathcal{L}(X_n)\to G_\epsilon$ for sufficiently small $a, b>0$.

Proof. By the Csizar inequality (Ref. 15),

$$\int q_b(x)\log(q_b(x)/p(x))\,dx\le(1/2)\left[\int|q_b(x)-p(x)|\,dx\right]^2.$$

The integral on the right is the total variation norm of the signed measure $p(x)\,dx-\eta_b(dx)$. The total variation norm topology on $P(R^d)$ is stronger than the Prohorov topology (Ref. 3, Chapter 2). Thus, the claim is immediate from Theorem 4.1 and the fact that $\eta_b\to\eta$ in $P(R^d)$ as $b\to 0$. □

Remark 4.1. Let $S=\text{support}(\eta)$. Since $\eta_b\to\eta$ in $P(R^d)$ for any open set $\hat{S}\subset P(R^d)$ containing S,

$$1\ge\liminf_{b\to 0}\eta_b(\hat{S})\ge\eta(\hat{S})=1.$$

Thus,

$$\lim_{b\to 0}\eta_b(\hat{S})=1.$$

By Corollary 4.1,

$$\lim_{b\to 0}\lim_{a\to 0}\limsup_{t\to\infty}P(X_n\in\hat{\bar{S}})=1,$$

for every open neighborhood $\hat{S}$ of S. Such a conclusion is of interest, e.g., when $h(\cdot)=-\nabla U(\cdot)$ as in the Gelfand-Mitter algorithm, with $S=$ the set of global minima of $U(\cdot)$. A more interesting result in this context would be to replace the triple limit above by a double limit, without $\lim_{b\to 0}$, but

with b specified as a function of a. This captures quantitatively the interplay between the effect of stepsize and the effect of noise. For the scalar Gelfand–Mitter algorithm, such a result is claimed in Ref. 16. More general results have been reported (Ref. 17).

Remark 4.2. Sometimes, it is of interest to consider a slightly more general version of (3), given by

$$X_{k+1} = X_k + a(h(X_k) + M_k) + \sqrt{a}b\sigma(X_k)W_k,$$

for a suitable $\sigma(\cdot) \in C(R^d; R^{d\times d})$. For example, in the global optimization algorithm of (1), one may want to increase the variance of noise in the regions where $U(\cdot)$ is relatively flat (i.e., its gradient is small in magnitude) and decrease it otherwise, in order to speed up the algorithm. One must then replace (4) by

$$dX(t) = h(X(t))\,dt + b\sigma(X(t))\,dW(t).$$

If $\sigma(\cdot)$ is Lipschitz and nondegenerate [i.e., the least eigenvalue of $\sigma(x)\sigma(x)^T$ is uniformly bounded away from zero], the foregoing analysis still goes through with only minor changes. Using the equality

$$E\left[\left(\int_0^t \langle Z(s), dW(s)\rangle\right)^2\right] = \int_0^t E[\|Z(s)\|^2]\,ds,$$

for appropriately defined $Z(\cdot)$, (7) can again be verified. The results of Refs. 10, 11, and 13, on which the rest of the argument hinges, continue to apply and, therefore, the conclusions remain unaltered.

References

1. GELFAND, S. B., and MITTER, S. K., *Recursive Stochastic Algorithms for Global Optimization in R^N*, SIAM Journal on Control and Optimization, Vol. 29, pp. 999–1018, 1991.
2. GELFAND, S. B., and MITTER, S. K., *Metropolis-Type Annealing Algorithm for Global Optimization in R^d*, SIAM Journal on Control and Optimization, Vol. 31, pp. 111–131, 1993.
3. BORKAR, V. S., *Probability Theory: An Advanced Course*, Springer Verlag, New York, New York, 1996.
4. HIRSCH, M. W., *Convergent Activation Dynamics in Continuous-Time Networks*, Neural Networks, Vol. 2, pp. 331–349, 1989.
5. LIPTSER, R. S., and SHIRYAYEV, A. N., *Statistics of Random Processes, I: General Theory*, Springer Verlag, New York, New York, 1977.

6. BHATTACHARYA, R. N., *Asymptotic Behavior of Several Dimensional Diffusions*, Stochastic Nonlinear Systems, Edited by L. Arnold and R. Lefever, Springer Verlag, Berlin, Germany, pp. 86–99, 1981.
7. BORKAR, V. S., and GHOSH, M. K., *Ergodic Control of Multidimensional Diffusions, II: Adaptive Control*, Applied Mathematics and Optimization, Vol. 21, pp. 191–220, 1990.
8. KUSHNER, H. J., *Approximation and Weak Convergence Methods for Random Processes*, MIT Press, Cambridge, Massachusetts, 1984.
9. MEYN, S. P., and TWEEDIE, R. L., *Markov Chains and Stochastic Stability*, Springer Verlag, New York, New York, 1994.
10. ARONSON, D. G., *Bounds for the Fundamental Solution of a Parabolic Equation*, Bulletin of the American Mathematical Society, Vol. 73, pp. 890–896, 1967.
11. BORKAR, V. S., *A Remark on the Attainable Distributions of Controlled Diffusions*, Stochastics, Vol. 18, pp. 17–23, 1986.
12. GŸONGI, I., *Mimicking the One-Dimensional Marginal Distribution of Processes Having an Itô Differential*, Probability Theory and Related Fields, Vol. 71, pp. 501–516, 1986.
13. LADYZENSKAJA, O. A., SOLONNIKOV, V. A., and URAL'CEVA, N. N., *Linear and Quasilinear Equations of Parabolic Type*, Translations of Mathematical Monographs, American Mathematical Society, Providence, Rhode Island, Vol. 23, 1968.
14. COVER, T. M., and THOMAS, J. A., *Elements of Information Theory*, John Wiley, New York, New York, 1991.
15. DEUSCHEL, J. D., and STROOCK, D., *Large Deviations*, Academic Press, New York, New York, 1989.
16. PFLUG, G., *Searching for the Best: Stochastic Approximation, Simulated Annealing, and Related Procedures*, Identification, Adaptation, Learning, Edited by S. Bittanti and G. Picci, Springer Verlag, Berlin, Germany, pp. 514–549, 1996.
17. GELFAND, S. B., Personal Communication, 1996.