Journal of Global Optimization **23:** 329–358, 2002.

Convergence Rates of Digital Diffusion Network Algorithms for Global Optimization with Applications to Image Estimation

G. YIN[1] and P.A. KELLY[2]
[1]*Department of Mathematics, Wayne State University, Detroit, MI 48202, USA (E-mail: gyin@math.wayne.edu)* [2]*Department of Electrical & Computer Engineering, University of Massachusetts, Amherst, MA 10013, USA (E-mail: kelly@ecs.umass.edu)*

Abstract. Motivated by the recent developments in digital diffusion networks, this work is devoted to the rates of convergence issue for a class of global optimization algorithms. By means of weak convergence methods, we show that a sequence of suitably scaled estimation errors converges weakly to a diffusion process (a solution of a stochastic differential equation). The scaling together with the stationary covariance of the limit diffusion process gives the desired rates of convergence. Application examples are also provided for some image estimation problems.

Key words: Global optimization, Recursive algorithm, Rate of convergence, Image estimation

1. Introduction

This work ascertains rates of convergence for a class of global optimization algorithms. The primary motivation of our study stems from applications to image estimation, namely, segmentation and restoration. Many of the image estimation problems can be recast to global optimization problems. For such problems, one often uses a simulated annealing-type algorithm to carry out the computation. Although the simulated annealing prevents the iterates from being trapped at local minima, straightforward sequential implementations can be too time-consuming for practical purposes. Recently, based on modifications of the Langevin algorithm and the Hopfield network, Wong (1991) suggested a diffusion network model. One of the main ideas in that paper is the use of parallel processors for the desired computation tasks, which allows speedup of the computation for optimization tasks.

Wong's approach can be outlined as follows. Let $\mathcal{E} : [0,1]^r \mapsto \mathbb{R}$ be an 'energy' function defined on the hypercube $[0,1]^r = [0,1] \times \cdots \times [0,1]$. Find the global minimizer of $\mathcal{E}(\cdot)$ by use of a neural network. Suppose that for all $t \geqslant 0$, and for each $\alpha = 1, \ldots, r$, $v_\alpha(t) \in [0,1]$ is the state at node α at time t and $v = (v_1, \ldots, v_r)^\tau \in [0,1]^r$ is an r-dimensional column vector (z^τ denotes the transpose of z). By injecting noise into a Hopfield network, one obtains the

dynamics. For $\alpha = 1, \ldots, r$, the system for the αth node is described by

$$\begin{aligned} v_\alpha(t) &= g(u_\alpha(t)), \\ du_\alpha(t) &= -\frac{\partial \mathcal{E}(v(t))}{\partial v_\alpha} dt + \widetilde{a}_\alpha(u(t)) dw_\alpha(t), \end{aligned} \tag{1.1}$$

where for $\alpha \leqslant r$, $\{w_\alpha(\cdot)\}$ are independent, standard and real-valued, Brownian motions, and $\widetilde{a}_\alpha(\cdot)$ and $g(\cdot)$ are appropriate functions. By choosing $\widetilde{a}_\alpha(u(t)) = [(2T)/g'(u_\alpha(t))]^{1/2}$, where g' denotes the derivative of g, $v(\cdot)$ becomes a stationary Markov process with stationary density

$$p_\infty(v) = (1/Z) \exp(-(1/T)\mathcal{E}(v)),$$

where Z is a normalizing factor

$$Z = \int \exp(-(1/T)\mathcal{E}(v)) dv \text{ so that } \int p_\infty(v) dv = 1.$$

Furthermore, by selecting $f(x) = g'(g^{-1}(x))$, for each $\alpha \leqslant r$,

$$dv_\alpha(t) = -f(v_\alpha(t)) \frac{\partial \mathcal{E}(v(t))}{\partial v_\alpha} dt + T f'(v_\alpha(t)) dt + \sqrt{2T f(v_\alpha(t))} dw_\alpha(t), \tag{1.2}$$

where T goes to zero sufficiently slowly.

Wong's diffusion machine is an analog one. One of the important features of his model is that by proper choice of $g(\cdot)$, $v(\cdot)$ is stationary. Moreover, although the process is defined on a bounded region, one need not worry about the reflecting boundaries, and need only consider diffusions instead of reflected diffusions. This is one of the most remarkable and significant contributions since it reduces much of the complexity and difficulty in dealing with the boundaries. Although Wong's work gives the desired diffusion equations, the rates of convergence were not considered. Subsequent work in this direction was carried out by Kesidis (1995). It was noted, however, that an analog implementation of the network does not appear to be practical for large-scale problems.

Problems arising from image estimation are frequently of large scale. For example, since a large number of pixels are involved and since each pixel is effectively represented by a component of a vector, even for a moderate sized segmentation problem, the dimension of the computation tasks can be very large. Taking advantages of Wong's diffusion network and overcoming the difficulties of the analog implementation, Cai et al. (1995) proposed a digital version of the diffusion network. The basic idea lies in using a stochastic difference equation in lieu of a continuous-time stochastic differential equation, which yields a discrete-time stochastic dynamic system. To obtain the asymptotic properties, a crucial step is to

show that the discrete iterates well approximate Wong's diffusion machine. Taking this into consideration, in our recent work Yin et al. (2000), we reveal the relationship of the discrete recursion and that of the continuous-time diffusion process by means of weak convergence methods and martingale averaging techniques. We have shown that by appropriate scaling, a suitably scaled sequence of the discrete iterates converges to Wong's analog diffusion machine. Nevertheless, the rates of convergence for such algorithms have not been considered. In this work, we continue the study initiated in Yin et al. (2000) by taking up the rates of convergence issues. Similar to that reference, we first deal with an algorithm defined on the entire $\mathbb{R}^r$. Subsequently, we restrict our attention to the r-dimensional hypercube $[0, 1]^r = [0, 1] \times \cdots \times [0, 1]$, which results in Wong's diffusion machine. We also examine variations of the algorithms including decreasing step size algorithms and algorithms with noisy measurements. In our numerical experiments, we show that our proposed numerical algorithms out perform the existing schemes. As shown in our numerical experiments, our recursive algorithms are more accurate than the existing procedures such as a maximum likelihood estimator. For example, in the first example considered in the Section 5 to follow, using maximum likelihood estimator, the error rate of approximation is 41%, whereas using our algorithm for image segmentation, the error rate is only 6.49%, a substantial improvement.

The rest of the paper is arranged as follows. Section 2 begins with the formulation of the algorithm. Section 3 proceeds with the rate of convergence analysis of the digital diffusion network algorithm. Unlike the global optimization algorithms treated in Yin (1999), the algorithm considered here uses constant step size ε and a restarting device. We show that a proper scaling (a continuous-time interpolation) of the discrete sequence converges to a diffusion process. The scaling factor together with the asymptotic covariance gives us the rates of convergence. Section 4 deals with the corresponding decreasing step-size algorithms, algorithms with additional contributing noise sources, and algorithms with iterates confined to $[0, 1]^r$. To illustrate the performance of the algorithm, Section 5 is devoted to a number of examples from image estimation. Finally, we close the paper with a few more remarks in Section 6. In the rest of the paper, for convenience, we use K to denote a generic positive constant with the convention $K + K = K$ and $KK = K$.

2. Recursive algorithm

Let $\mathcal{E}(\cdot) : \mathbb{R}^r \mapsto \mathbb{R}$, $f(\cdot) : \mathbb{R} \mapsto \mathbb{R}$, and $g(\cdot) : \mathbb{R} \mapsto \mathbb{R}$. In what follows, unless otherwise noted, a Greek letter α or β denotes a component (or corresponding to a processor) and ι, k and n denote the indices of the iterations. Without loss of generality, we assume that each processor in the diffusion network controls one component of the underlying vector. The treatment for the case that one processor controls more than one components is the same. Inspired by the ideas in stochastic approximation (see, for example, the up-to-date treatment of Kushner and Yin (1997)) and the simulated annealing algorithm Gelfand and Mitter (1991), we pro-

posed the following recursive algorithm with a periodic restarting device in Yin et al. (2000). The idea is to partially reset the step-size sequences once a while, which allows us to obtain the desired limit continuous-time diffusion easily. Meanwhile, it is easily implementable and does not add any more complexity in the computation. In the actual implementation of stochastic approximation algorithms, very often one wishes to use a constant step size in lieu of a sequence of decreasing step sizes. Constant step size algorithms have advantages in that, first, they are easily implementable, and second, such algorithms often provide better tracking properties. For each $\iota \geqslant 0$ and for a sequence $\{\varpi_k^{\iota n}\}$ (in either $\mathbb{R}^r$ or $\mathbb{R}$), $\varpi_k^{\iota n} = \varpi_{\iota n+k}$. For $\alpha \leqslant r$, the algorithm for the diffusion network, which periodically (with period n) resets the step-size sequences, takes the form: For $0 \leqslant k < n$,

$$v_{\alpha,k+1}^{\iota n} = v_{\alpha,k}^{\iota n} - a_k^{\iota n} f(v_{\alpha,k}^{\iota n}) \frac{\partial \mathcal{E}(v_k^{\iota n})}{\partial v_\alpha} + c_k^{\iota n} f'(v_{\alpha,k}^{\iota n}) + b_k^{\iota n} \sqrt{f(v_{\alpha,k}^{\iota n})} W_{\alpha,k}^{\iota n}, \quad (2.1)$$

where $\varepsilon > 0$ is a small parameter, $A_0 > 1$, and

$$\begin{aligned} a_k^{\iota n} &= \varepsilon, \\ b_k^{\iota n} &= \sqrt{2\varepsilon}/[\sqrt{\ln[\varepsilon k + A_0)}], \\ c_k^{\iota n} &= \varepsilon/\ln[\varepsilon k + A_0]. \end{aligned} \quad (2.2)$$

In Yin et al. (2000), we proved the convergence of the algorithm by first obtaining the weak convergence of an interpolated sequence of the iterates to the stochastic differential equations

$$\begin{aligned} dv_\alpha(t) = &- f(v_\alpha(t)) \frac{\partial \mathcal{E}(v(t))}{\partial v_\alpha} dt + \frac{1}{\ln(t + A_0)} f'(v_\alpha(t)) dt \\ &+ \sqrt{\frac{2f(v_\alpha(t))}{\ln(t + A_0)}} dw_\alpha(t), \ \alpha = 1, \ldots, r. \end{aligned} \quad (2.3)$$

We then used a result of Chiang et al. (1987) and established the desired convergence. Note that in the case of $v \in [0, 1]^r$, (2.3) is precisely Wong's diffusion machine with $T(t) = 1/\ln(t + A_0)$. In this work, we continue our investigation by concentrating on the rates of convergence of the algorithm. In the analysis to follow, it is often more convenient to work with vector-valued processes. We thus use the following notation:

$$\begin{aligned} W_k^n &= (W_{1,k}^n, \ldots, W_{r,k}^n)^\tau, \\ F(v) &= \operatorname{diag}(f(v_1), \ldots, f(v_r)), \\ D(v) &= (f'(v_1), \ldots, f'(v_r))^\tau, \\ \Phi(v) &= \operatorname{diag}(\sqrt{f(v_1)}, \ldots, \sqrt{f(v_r)}), \\ \mathcal{E}_v(v) &= \nabla_v \mathcal{E}(v) = \left(\frac{\partial \mathcal{E}(v)}{\partial v_1}, \ldots, \frac{\partial \mathcal{E}(v)}{\partial v_r} \right)^\tau, \end{aligned}$$

where $\mathrm{diag}(d_1, \ldots, d_r)$ denotes a diagonal matrix with diagonal entries d_1 through d_r. As in Yin et al. (2000), for the analysis to follow, we fix ι. Without loss of generality, take $\iota = 1$ henceforth. In view of the vector notation (with $\iota = 1$), (2.1) and (2.3) can be written as

$$v_{k+1}^n = v_k^n - \varepsilon F(v_k^n)\mathcal{E}_v(v_k^n) + c_k^n D(v_k^n) + b_k^n \Phi(v_k^n) W_k^n, \tag{2.4}$$

and

$$dv(t) = -F(v(t))\nabla\mathcal{E}(v(t))dt + \frac{1}{\ln(t+A_0)}D(v(t))dt + \sqrt{\frac{2f(v(t))}{\ln(t+A_0)}}dw(t), \tag{2.5}$$

respectively. The analysis in Yin et al. (2000) lies upon an interpolated process $v^\varepsilon(\cdot)$ defined by $v^\varepsilon(t) = v_k^n$ for $t \in [\varepsilon k, \varepsilon k + \varepsilon)$. We have shown that $v^\varepsilon(\cdot)$ converges weakly to the solution of (2.5).

REMARK 2.1. In (2.1), the step-size sequence $\{a_k^n\}$ is a constant ε. In what follows, for convenience and with a slight abuse of terminology, we refer to this algorithm as a constant-step-size algorithm. If a_k^n is a decreasing sequence, we refer to the corresponding algorithm as a decreasing-step-size algorithm. There should be no confusion from the context.

To proceed, let us make the following assumptions:

(**A1**) The functions $f(\cdot)$ and $g(\cdot)$ are continuously differentiable, and $\mathcal{E}_{vv}$ (the second partial derivative of the function $\mathcal{E}(\cdot)$) exists and is continuous and bounded such that
- (a) $\mathcal{E}(v) \geqslant 0$ for all $v \in \mathbb{R}^r$, $\min_v \mathcal{E}(v) = 0$, and the set $\mathsf{M} = \{v \in \mathbb{R}^r; \mathcal{E}_v(v) = 0\}$ consists of finitely many isolated points, and there is a global minimizer $v^* \in \mathbb{R}^r$;
- (b) $f(z) \geqslant 0$ for all $z \in \mathbb{R}$, and $f(\cdot)$ is bounded with bounded derivative;
- (c) the inverse $g^{-1}(\cdot)$ exists and is continuously differentiable;
- (d) $f(z) = g'(g^{-1}(z))$;
- (e) $v^\varepsilon(\varepsilon t_\varepsilon + \cdot)$ converges weakly to v^*, where t_ε is a sequence of real numbers satisfying $t_\varepsilon \geqslant 0$ and $\varepsilon t_\varepsilon \to \infty$ as $\varepsilon \to 0$.

(**A2**) For each $\alpha \leqslant r$, $\{W_{\alpha,k}^n\}$ is a sequence of independent and identically distributed random variables such that
- (a) $EW_{\alpha,k}^n = 0$;
- (b) $E(W_{\alpha,k}^n)^2 = 1$;
- (c) for $\alpha \neq \beta$, $W_{\alpha,k}^n$ and $W_{\beta,k}^n$ are independent.

(**A3**) There is a twice continuously differentiable Liapunov function $U(\cdot) : \mathbb{R}^r \mapsto \mathbb{R}$ for (2.3) such that
- (a) $U(v) \geqslant 0$ for all v, $U(v) \to \infty$ as $|v| \to \infty$;
- (b) $|U_v(v)| \leqslant K(1+U^{1/2}(v))$, $|U_{vv}(\cdot)| \leqslant K$, and $|\mathcal{E}_v(v)| \leqslant K(1+U^{1/2}(v))$;
- (c) $U_v^\tau(v)F(v)\mathcal{E}_v(v) \geqslant \widetilde{\lambda}U(v)$ for some $\widetilde{\lambda} > 0$ and $v \notin \mathsf{M}$.

REMARK 2.2. As was mentioned in the Introduction, for more generality, we first consider the case $\mathcal{E}(\cdot)$ being defined on $\mathbb{R}^r$. Later $\mathcal{E}(\cdot)$ defined on the hypercube $[0,1]^r$ is treated as a special case. Note also that the condition on $\{W_k^n\}$ indicates that the perturbing noise decouples among different processors, a property is helpful in the analysis and actual computing. Since the noise sequence $\{W_k^n\}$ is added by us, it is at our disposal. Thus, it is more convenient to use a sequence of uncorrelated random variables.

3. Rates of convergence

3.1. ERROR BOUNDS

The purpose of this section is to establish error bounds of the scaled sequences of estimation errors. We prove the desired bounds by using a Liapunov function.

THEOREM 3.1. *Suppose* (A1)–(A3) *are satisfied. Then for* $\{v_k^n\}$ *defined by* (2.1) *and for sufficiently large* k,

$$EU(v_k^n) = O(1), \tag{3.1}$$

and

$$\ln(\varepsilon k + A_0)EU(v_k^n) = O(1). \tag{3.2}$$

Proof. Although generally $v^* \neq 0$, for notational simplicity, we assume that $v^* = 0$ in the proof of this theorem. We first prove (3.1). Denote the conditional expectation with respect to the σ-algebra generated by $\{W_j^n;\, j < k\}$ by E_k^n. Using the Liapunov function $U(\cdot)$ and $E_k^n W_k^n = 0$, we have

$$\begin{aligned}
&E_k^n U(v_{k+1}^n) - U(v_k^n)\\
&\quad \leqslant U_v^\tau(v_k^n)\left(-\varepsilon F(v_k^n)\mathcal{E}_v(v_k^n) + \frac{\varepsilon}{\ln(\varepsilon k + A_0)}D(v_k^n)\right)\\
&\quad + K\varepsilon^2\left(F(v_k^n)\mathcal{E}_v(v_k^n)\right)^\tau\left(\int_0^1 U_{vv}(v_k^n + s(v_{k+1}^n - v_k^n))ds\right)\left(F(v_k^n)\mathcal{E}_v(v_k^n)\right)\\
&\quad + \frac{K\varepsilon^2 D^\tau(v_k^n)}{[\ln(\varepsilon k + A_0)]^2}\left(\int_0^1 U_{vv}(v_k^n + s(v_{k+1}^n - v_k^n))ds\right)D(v_k^n)\\
&\quad + \frac{K\varepsilon}{\ln(\varepsilon k + A_0)}\mathrm{tr}\Bigg(\int_0^1 U_{vv}(v_k^n + s(v_{k+1}^n - v_k^n))ds\\
&\quad \times [E_k^n W_k^n W_k^{n,\tau}][\Phi^\tau(v_k^n)\Phi(v_k^n)]\Bigg).
\end{aligned} \tag{3.3}$$

In view of (A2),

$$E_k^n W_k^n W_k^{n,\tau} = I, \text{ (the identity matrix).}$$

Thus the boundedness of $f(\cdot)$ and $U_{vv}(\cdot)$ implies that

$$\left|\frac{2\varepsilon}{\ln(\varepsilon k+A_0)}\mathrm{tr}\Big(\int_0^1 U_{vv}(v_k^n+s(v_{k+1}^n-v_k^n))ds \times [E_k^n W_k^n W_k^{n,\tau}[\Phi^\tau(v_k^n)\Phi(v_k^n)]\Big)\right| \leqslant K\frac{\varepsilon}{\ln(\varepsilon k+A_0)}, \tag{3.4}$$

where $\mathrm{tr}(A)$ denotes the trace of A. The boundedness of $f'(\cdot)$ yields that

$$\left|\frac{\varepsilon^2 D^\tau(v_k^n)}{[\ln(\varepsilon k+A_0)]^2}\left(\int_0^1 U_{vv}(v_k^n+s(v_{k+1}^n-v_k^n))ds\right)D(v_k^n)\right| \leqslant K\frac{\varepsilon^2}{[\ln(\varepsilon k+A_0)]^2}. \tag{3.5}$$

Moreover, we also have

$$\left|\varepsilon^2\left(F(v_k^n)\mathcal{E}_v(v_k^n)\right)^\tau\left(\int_0^1 U_{vv}(v_k^n+s(v_{k+1}^n-v_k^n))ds\right)\left(F(v_k^n)\mathcal{E}_v(v_k^n)\right)\right| \leqslant K\varepsilon^2(1+U(v_k^n)). \tag{3.6}$$

Combining (3.4)–(3.6), the last three terms in (3.3) are bounded by

$$O(\varepsilon^2)U(v_k^n)+O\left(\varepsilon^2+\frac{\varepsilon^2}{[\ln(\varepsilon k+A_0)]^2}+\frac{\varepsilon}{\ln(\varepsilon K+A_0)}\right) = O(\varepsilon^2)U(v_k^n)+O\left(\frac{\varepsilon}{\ln(\varepsilon K+A_0)}\right). \tag{3.7}$$

If $v_k^n \notin \mathsf{M}$, $U_v^\tau(v_k^n)F(v_k^n)\mathcal{E}_v(v_k^n) \geqslant \widetilde{\lambda}U(v_k^n)$ by (A3) (c). Consequently, whenever $v_k^n \notin \mathsf{M}$, there is some $\lambda > 0$,

$$-U_v^\tau(v_k^n)F(v_k^n)\mathcal{E}_v(v_k^n) \leqslant -\lambda U(v_k^n). \tag{3.8}$$

It follows from (A1) (a), there is a constant vector $\widetilde{e} \in \mathbb{R}^r$ such that if $v_k^n \in \mathsf{M}$, $v_k^n+\varepsilon\widetilde{e} \notin \mathsf{M}$. As a result, by (A3) (c),

$$\begin{aligned}
&-U_v^\tau(v_k^n)F(v_k^n)\mathcal{E}_v(v_k^n)\\
&\quad=-U_v^\tau(v_k^n)F(v_k^n)\mathcal{E}_v(v_k^n)I_{\{v_k^n\notin\mathsf{M}\}}-U_v^\tau(v_k^n)F(v_k^n)\mathcal{E}_v(v_k^n)I_{\{v_k^n\in\mathsf{M}\}}\\
&\quad\leqslant-\lambda U(v_k^n)I_{\{v_k^n\notin\mathsf{M}\}}-U_v^\tau(v_k^n+\varepsilon\widetilde{e})F(v_k^n+\varepsilon\widetilde{e})\mathcal{E}_v(v_k^n+\varepsilon\widetilde{e})I_{\{v_k^n\in\mathsf{M}\}}\\
&\qquad+U_v^\tau(v_k^n+\varepsilon\widetilde{e})F(v_k^n+\varepsilon\widetilde{e})\mathcal{E}_v(v_k^n+\varepsilon\widetilde{e})I_{\{v_k^n\in\mathsf{M}\}}.
\end{aligned} \tag{3.9}$$

In addition,

$$\begin{aligned}
&-U_v^\tau(v_k^n+\varepsilon\widetilde{e})F(v_k^n+\varepsilon\widetilde{e})\mathcal{E}_v(v_k^n+\varepsilon\widetilde{e})I_{\{v_k^n\in\mathsf{M}\}}\\
&\quad\leqslant-\lambda U(v_k^n+\varepsilon\widetilde{e})I_{\{v_k^n\in\mathsf{M}\}}\\
&\quad\leqslant-\lambda U(v_k^n)I_{\{v_k^n\in\mathsf{M}\}}+\left|\varepsilon\widetilde{e}\left(\int_0^1 U_v^\tau(v_k^n+s\varepsilon\widetilde{e})ds\right)I_{\{v_k^n\in\mathsf{M}\}}\right|\\
&\quad\leqslant-\lambda U(v_k^n)I_{\{v_k^n\in\mathsf{M}\}}+O(\varepsilon)(1+U(v_k^n)),
\end{aligned} \tag{3.10}$$

and for the last term in (3.9),

$$\begin{aligned}
&\left|U_v^\tau(v_k^n+\varepsilon\widetilde{e})F(v_k^n+\varepsilon\widetilde{e})\mathcal{E}_v(v_k^n+\varepsilon\widetilde{e})I_{\{v_k^n\in\mathsf{M}\}}\right| \\
&\quad=\left|\varepsilon\widetilde{e}\left(\int_0^1 U_v^\tau(v_k^n+s\varepsilon\widetilde{e})F(v_k^n+\varepsilon\widetilde{e})\mathcal{E}_v(v_k^n+\varepsilon\widetilde{e})ds\right)_v I_{\{v_k^n\in\mathsf{M}\}}\right| \\
&\quad\leqslant O(\varepsilon)(1+U(v_k^n)).
\end{aligned}\tag{3.11}$$

Thus (3.8), (3.9), (3.10), and (3.11) lead to

$$-U_v^\tau(v_k^n)F(v_k^n)\mathcal{E}_v(v_k^n)\leqslant-\lambda U(v_k^n)+O(\varepsilon)(1+U(v_k^n)).$$

Finally,

$$\left|\frac{\varepsilon}{\ln(\varepsilon k+A_0)}U_v^\tau(v_k^n)D(v_k^n)\right|\leqslant K\frac{\varepsilon}{\ln(\varepsilon k+A_0)}(1+U(v_k^n)).\tag{3.12}$$

Using (3.3)–(3.12), we arrive at

$$\begin{aligned}
&E_k^nU(v_{k+1}^n) \\
&\quad\leqslant(1-\lambda\varepsilon)U(v_k^n)+O(\varepsilon^2)U(v_k^n)+O\left(\frac{\varepsilon}{\ln(\varepsilon k+A_0)}\right)(1+U(v_k^n)) \\
&\quad\leqslant(1-\lambda_0\varepsilon)U(v_k^n)+O\left(\frac{\varepsilon}{\ln(\varepsilon k+A_0)}\right)EU(v_k^n)+O\left(\frac{\varepsilon}{\ln(\varepsilon k+A_0)}\right),
\end{aligned}$$

where $0<\lambda_0<\lambda$. Iterating on the above inequality and taking expectation, we obtain

$$\begin{aligned}
EU(v_{k+1}^n)\leqslant&(1-\lambda_0\varepsilon)^kEU(v_0^n) \\
&+K\sum_{j=0}^k\frac{\varepsilon(1-\lambda_0\varepsilon)^{k-j}}{\ln(\varepsilon j+A_0)}+K\sum_{j=0}^k\frac{\varepsilon(1-\lambda_0\varepsilon)^{k-j}}{\ln(\varepsilon j+A_0)}EU(v_j^n).
\end{aligned}\tag{3.13}$$

Denoting

$$\mu_k^n=(1-\lambda_0\varepsilon)^kEU(v_0^n)+K\sum_{j=0}^k\frac{\varepsilon}{\ln(\varepsilon j+A_0)}(1-\lambda_0\varepsilon)^{k-j},$$

the Gronwall inequality then yields

$$EU(v_{k+1}^n)\leqslant\mu_k^n\exp\left(K\sum_{j=0}^k\frac{\varepsilon}{\ln(\varepsilon j+A_0)}(1-\lambda_0\varepsilon)^{k-j}\right),\tag{3.14}$$

and hence (3.1) follows.

Next, we prove (3.2). In view of (3.14), we have

$$\ln(\varepsilon k + A_0) EU(v_{k+1}^n)$$
$$\leqslant \ln(\varepsilon k + A_0)\mu_k^n \exp\left(K \sum_{j=0}^{k} \frac{\varepsilon}{\ln(\varepsilon j + A_0)} (1 - \lambda_0 \varepsilon)^{k-j} \right).$$

As a result, to obtain the desired bound, it suffices to show $\ln(\varepsilon k + A_0)\mu_k^n = O(1)$. To this end, it suffices to show

$$\ln(\varepsilon k + A_0)(1 - \lambda_0 \varepsilon)^k EU(v_0^n) = O(1)$$
$$\ln(\varepsilon k + A_0) \sum_{j=0}^{k} \frac{\varepsilon}{\ln(\varepsilon j + A_0)} (1 - \lambda_0 \varepsilon)^{k-j} = O(1). \tag{3.15}$$

The first inequality in (3.15) is easily verified by observing the exponential decay property of $(1 - \lambda_0 \varepsilon)^k$. As for the second one, using a summation by parts,

$$\ln(\varepsilon k + A_0) \sum_{j=0}^{k} \frac{\varepsilon}{\ln(\varepsilon j + A_0)} (1 - \lambda_0 \varepsilon)^{k-j}$$
$$= \sum_{j=0}^{k} \varepsilon (1 - \lambda_0 \varepsilon)^{k-j} + \ln(\varepsilon k + A_0)$$
$$\times \sum_{j=0}^{k-1} \left(\frac{1}{\ln(\varepsilon j + A_0)} - \frac{1}{\ln(\varepsilon (j+1) + A_0)} \right) \sum_{i=0}^{j} \varepsilon (1 - \lambda_0)^{k-i}$$
$$\leqslant O(1) + K \sum_{j=0}^{k-1} \frac{\ln\left(1 + \frac{1}{j}\right)}{(\ln(\varepsilon j + A_0))^2} \sum_{i=0}^{j} \varepsilon (1 - \lambda_0 \varepsilon)^{k-i}. \tag{3.16}$$

Upon using

$$\ln(1 + x) \leqslant x \text{ for any real } x > 0, \text{ and } \sum_{i=0}^{j} (1 - \lambda_0 \varepsilon)^{k-i} \leqslant j (1 - \lambda_0 \varepsilon)^{k-j},$$

(3.16) yields that for some κ_0 with $0 < \kappa_0 < 1$,

$$\ln(\varepsilon k + A_0) \sum_{j=0}^{k} \frac{\varepsilon}{\ln(\varepsilon k + A_0)} (1 - \lambda_0 \varepsilon)^{k-j}$$
$$\leqslant O(1) + \kappa_0 \ln(\varepsilon k + A_0) \sum_{j=0}^{k} \frac{\varepsilon (1 - \lambda_0 \varepsilon)^{k-j}}{\ln(\varepsilon j + A_0)},$$

which in turn leads to

$$(1-\kappa_0)\ln(\varepsilon k+A_0)\sum_{j=0}^{k}\frac{\varepsilon(1-\lambda_0\varepsilon)^{k-j}}{\ln(\varepsilon j+A_0)}\leqslant O(1). \tag{3.17}$$

Finally,

$$\ln(\varepsilon(k+1)+A_0)EU(v_{k+1}^n)=\frac{\ln(\varepsilon(k+1)+A_0)}{\ln(\varepsilon k+A_0)}\left(\ln(\varepsilon k+A_0)EU(v_{k+1}^n)\right).$$

The boundedness of $\ln(\varepsilon(k+1)+A_0)/\ln(\varepsilon k+A_0)$ and (3.17) then yield (3.2). □

3.2. NOTION OF WEAK CONVERGENCE

To obtain further limit results, we use the methods of weak convergence. The notion of weak convergence of probability measures is a substantial extension of convergence in distribution, and is a powerful machinery for a wide range of applications. For completeness and reference, we mention some of the basic definitions and notation for weak convergence. Let X_n and X be $\mathbb{R}^r$-valued random variables. We say that X_n converges weakly to X iff for any bounded and continuous function $\psi(\cdot)$, $E\psi(X_n)\to E\psi(X)$. Similar to the notion of compactness, we often wish to say that 'no probabilities are lost' for a sequence. Such a notion is referred to as tightness. The sequence $\{X_n\}$ is tight iff for each $\eta>0$, there is a compact set K_η such that $P(X_n\in K_\eta)\geqslant 1-\eta$ for all n. The definitions of weak convergence and tightness extend to random variables in a metric space. On a complete separable metric space, tightness is equivalent to relative compactness. This is known as Prohorov's Theorem. By using this theorem, we are able to extract convergent subsequences once tightness is verified. In the weak convergence analysis, it is more convenient to work with $D^r[0,\infty)$, the space of functions that are right continuous, have left limits, endowed with the Skorohod topology (Ethier and Kurtz, 1986, Kushner and Yin, 1997). Moreover, for convenience, we often use a device known as Skorohod representation. Suppose that X_n converges weakly to X. Then enlarging the probability space if necessary, the Skorohod representation allows us to find $\widetilde{X}_n$ and $\widetilde{X}$ such that they have the same distribution as that of X_n and X, respectively, and that $\widetilde{X}_n\to\widetilde{X}$ w.p.1. In what follows, when we use such a device, without loss of generality and for notational simplicity, we will not use the tilde symbol. The application of weak convergence methods usually requires first tightness be proved and then the limit process be characterized.

Define

$$v_k^n=\sqrt{\ln(\varepsilon k+A_0)}[v_k^n-v^*]. \tag{3.18}$$

The proof presented in Section 3.1 leads to:

COROLLARY 3.2. *Assume the conditions of Theorem 3.1. If the Liapunov function $U(\cdot)$ is locally (near v^*) quadratic, then there is a sequence of real numbers, $\{M_\varepsilon\}$, satisfying $M_\varepsilon \geqslant t_\varepsilon$ such that $\{(v_k^n - v^*),\ k \geqslant M_\varepsilon\}$ is tight or bounded in probability. That is for any $\eta > 0$, there exists a K_η such that*

$$P(|v_k^n - v^*| \geqslant K_\eta) < \eta \ \text{ for all } k \geqslant M_\varepsilon.$$

REMARK 3.3. In what follows, we work with the sequence $\{v_{k+M_\varepsilon}^n - v^*\}$. Thus all the interpolations etc. should be defined for the indices $k + M_\varepsilon$ etc. Nevertheless, for notational simplicity, we shall still write v_k^n throughout.

Define a piecewise constant interpolation of ν_k^n as

$$\nu^\varepsilon(t) = \nu_k^n \ \text{ for } \ t \in [\varepsilon k, \varepsilon(k+1)).$$

In view of 3.3, what we are really working with is

$$\nu^\varepsilon(t) = \nu_{k+M_\varepsilon}^n \ \text{ for } \ t \in [\varepsilon(k+M_\varepsilon), \varepsilon(k+M_\varepsilon)+\varepsilon).$$

Nevertheless, the notation without M_ε is much simpler as can be seen in what follows.

3.3. DIFFUSION LIMIT

Note that $\mathcal{E}(v^*) = 0$. Using (2.4) with $\iota = 1$, define

$$H = (\partial/\partial v)(F(v^*)\mathcal{E}_v(v^*)), \ \text{ and } \ \widetilde{v}_k^n = v_k^n - v^*. \tag{3.19}$$

We linearize the recursion about v^*. It leads to

$$\begin{aligned}\widetilde{v}_{k+1}^n =& \widetilde{v}_k^n - \varepsilon H\widetilde{v}_k^n + c_k^n D(v^*) + c_k^n[D(v_k^n) - D(v^*)] + b_k^n \Phi(v^*)W_k^n \\ &+ b_k^n[\Phi(v_k^n) - \Phi(v^*)]W_k^n + o(\varepsilon)O\left(|\widetilde{v}_k^n|\right),\end{aligned}$$

where the last term comes from the remainder in the Taylor expansion. Using the definition (3.18), the above equation can be written as

$$\begin{aligned}v_{k+1}^n =& \sqrt{\frac{\ln(\varepsilon(k+1)+A_0)}{\ln(\varepsilon k + A_0)}}\nu_k^n - \varepsilon\sqrt{\frac{\ln(\varepsilon(k+1)+A_0)}{\ln(\varepsilon k + A_0)}}H\nu_k^n \\ &+ \frac{\varepsilon}{\sqrt{\ln(\varepsilon k + A_0)}}\sqrt{\frac{\ln(\varepsilon(k+1)+A_0)}{\ln(\varepsilon k + A_0)}}D(v^*) \\ &+ \sqrt{2\varepsilon}\sqrt{\frac{\ln(\varepsilon(k+1)+A_0)}{\ln(\varepsilon k + A_0)}}\Phi(v^*)W_k^n \\ &+ \frac{\varepsilon}{\sqrt{\ln(\varepsilon k + A_0)}}\sqrt{\frac{\ln(\varepsilon(k+1)+A_0)}{\ln(\varepsilon k + A_0)}}[D(v_k^n) - D(v^*)]\end{aligned}$$

$$+\sqrt{2\varepsilon}\sqrt{\frac{\ln(\varepsilon(k+1)+A_0)}{\ln(\varepsilon k+A_0)}}[\Phi(v_k^n)-\Phi(v^*)]W_k^n$$
$$+\sqrt{\frac{\ln(\varepsilon(k+1)+A_0)}{\ln(\varepsilon k+A_0)}}o(\varepsilon)O\left(\varepsilon|v_k^n|\right). \tag{3.20}$$

To carry out the analysis, use a truncation device (Kushner and Yin, 1997, p. 248). That is, for an arbitrary $N>0$, denote $S_N=\{x;|x|\leqslant N\}$ and let $v^{\varepsilon,N}(\cdot)$ be the process that is equal to $v^\varepsilon(\cdot)$ up until the first exit from S_N and that

$$\lim_{K\to\infty}\limsup_{\varepsilon\to 0}P(\sup_{t\leqslant T}|v^{\varepsilon,N}(t)|\geqslant K)=0,\quad \text{for each } T<\infty \text{ and } N<\infty.$$

Let a truncation function $q_N(\cdot)$ be defined as

$$q_N(x)=\begin{cases}1, & |x|\leqslant N,\\ 0, & |x|>N+1.\end{cases}$$

In lieu of (3.20), we work with a truncated process. The corresponding recursion is given by

$$\begin{aligned}v_{k+1}^{n,N}=&\sqrt{\frac{\ln(\varepsilon(k+1)+A_0)}{\ln(\varepsilon k+A_0)}}v_k^{n,N}-\varepsilon\sqrt{\frac{\ln(\varepsilon(k+1)+A_0)}{\ln(\varepsilon k+A_0)}}Hv_k^{n,N}\\
&+\frac{\varepsilon}{\sqrt{\ln(\varepsilon k+A_0)}}\sqrt{\frac{\ln(\varepsilon(k+1)+A_0)}{\ln(\varepsilon k+A_0)}}D(v^*)\\
&+\sqrt{2\varepsilon}\sqrt{\frac{\ln(\varepsilon(k+1)+A_0)}{\ln(\varepsilon k+A_0)}}\Phi(v^*)W_k^n\\
&+\frac{\varepsilon}{\sqrt{\ln(\varepsilon k+A_0)}}\sqrt{\frac{\ln(\varepsilon(k+1)+A_0)}{\ln(\varepsilon k+A_0)}}[D(v_k^{n,N})-D(v^*)]q_N(v_k^{n,N})\\
&+\sqrt{2\varepsilon}\sqrt{\frac{\ln(\varepsilon(k+1)+A_0)}{\ln(\varepsilon k+A_0)}}[\Phi(v_k^{n,N})-\Phi(v^*)]q_N(v_k^{n,N})W_k^n\\
&+\sqrt{\frac{\ln(\varepsilon(k+1)+A_0)}{\ln(\varepsilon k+A_0)}}o(\varepsilon)O\left(|v_k^{n,N}|\right)q_N(v_k^{n,N}).\end{aligned} \tag{3.21}$$

In the above, we have also used the truncation $v_k^{n,N}$ for v_k^n.

REMARK 3.4. Note that for each N, $\{v_k^{n,N}\}$ and $\{v_k^{n,N}\}$ are uniformly bounded. That is, the bound may depend on N, but it is independent of k. This fact will be used crucially in what follows.

Our plan is: We first establish the tightness of $\{\nu^{\varepsilon,N}(\cdot)\}$. Then we work with $\{\nu^{\varepsilon,N}(\cdot)\}$, and derive its weak convergence. Finally, we let $N \to \infty$ to conclude the proof.

3.3.1. *Tightness of* $\{\nu^{\varepsilon,N}(\cdot)\}$

Note that $\nu^{\varepsilon,N}(\cdot)$ is in $D^r[0,\infty)$. We state a theorem that gives the tightness of this sequence of functions.

THEOREM 3.5. *Assume the conditions of Corollary 3.2 hold. Then* $\{\nu^{\varepsilon,N}(\cdot)\}$ *is tight in* $D^r[0,\infty)$.

Proof. We use the tightness criteria due to Kurtz (see Ethier and Kurtz (1986, p.137), and Kushner (1984, p.47) to obtain the desired result. Use E_t^ε and E_k^n to denote the conditional expectations with respect to the σ-algebras generated by $\{\nu^{\varepsilon,N}(u),\ u \leqslant t\}$ and $\{\nu_0^{\varepsilon,N}, W_j^n,\ j < k\}$, respectively.

Note that

$$\sqrt{\frac{\ln(\varepsilon(k+1)+A_0)}{\ln(\varepsilon k + A_0)}} = 1 + O\left(\frac{\ln\left(1+\varepsilon/(\varepsilon k + A_0)\right)}{\ln(\varepsilon k + A_0)}\right). \tag{3.22}$$

Thus (3.21) can be written as

$$\begin{aligned}
\nu_{k+1}^{n,N} =& \nu_k^{n,N} - \varepsilon H \nu_k^{n,N} + \frac{\varepsilon}{\sqrt{\ln(\varepsilon k + A_0)}} D(v^*) + \sqrt{2\varepsilon}\Phi(v^*) W_k^n \\
&+ \frac{\varepsilon}{\sqrt{\ln(\varepsilon k + A_0)}} [D(v_k^{n,N}) - D(v^*)] q_N(\nu_k^{n,N}) \\
&+ \sqrt{2\varepsilon}[\Phi(v_k^{n,N}) - \Phi(v^*)] q_N(\nu_k^{n,N}) W_k^n \\
&+ o(\varepsilon) O\left(|\nu_k^{n,N}|\right) q_N(\nu_k^{n,N}) + e_n.
\end{aligned} \tag{3.23}$$

Comparing (3.20) with (3.23), we see that e_n collects all terms involving

$$O\left(\frac{\ln\left(1+\varepsilon/(\varepsilon k + A_0)\right)}{\ln(\varepsilon k + A_0)}\right)$$

due to the expansion (3.22) so it is asymptotically unimportant in the sense that

$$\sum_{j=\lfloor t/\varepsilon \rfloor}^{\lfloor (t+s)/\varepsilon \rfloor - 1} e_j \to 0 \ \text{ in probability uniformly in } \ t. \tag{3.24}$$

In the above, $\lfloor z \rfloor$ denotes the integer part of $z \in \mathbb{R}$. However, for simplicity, in what follows, we will not use the $\lfloor \cdot \rfloor$ notation.

Using (3.23) and the interpolation,

$$\begin{aligned}
\nu^{\varepsilon,N}(t+s)-\nu^{\varepsilon,N}(t) &= -\varepsilon\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1} H\nu_j^{n,N} + \sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}\frac{\varepsilon}{\sqrt{\ln(\varepsilon j+A_0)}}D(v^*)\\
&+\sqrt{2\varepsilon}\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}\Phi(v^*)W_j^n + \sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1} o(\varepsilon)O\left(|\nu_j^{n,N}|\right)q_N(\nu_j^{n,N})\\
&+\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}\frac{\varepsilon}{\sqrt{\ln(\varepsilon j+A_0)}}[D(v_j^{n,N})-D(v^*)]q_N(\nu_j^{n,N})\\
&+\sqrt{2\varepsilon}\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}[\Phi(v_j^{n,N})-\Phi(v^*)]q_N(\nu_j^{n,N})W_j^n + \sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1} e_j. \qquad (3.25)
\end{aligned}$$

Since the last term above goes to 0 in probability by virtue of (3.24), we shall disregard it in what follows and concentrate on the rest of the terms only.

By virtue of the boundedness of $\nu_j^{n,N}$, for s sufficiently small,

$$\begin{aligned}
E_t^\varepsilon\left|\varepsilon\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1} H\nu_j^{n,N}\right|^2 &\leqslant \varepsilon^2\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} E_t^\varepsilon|H|^2|\nu_j^{n,N}|^2\\
&\leqslant K\left(\frac{t+s}{\varepsilon}-\frac{t}{\varepsilon}\right)^2\varepsilon^2 = Ks^2 \leqslant Ks.
\end{aligned}$$

Similarly,

$$E_t^\varepsilon\left|\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}\frac{\varepsilon}{\sqrt{\ln(\varepsilon j+A_0)}}[D(v_j^{n,N})-D(v^*)]q_N(\nu_j^{n,N})\right|^2 \leqslant Ks,$$

and

$$E_t^\varepsilon\left|\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1} o(\varepsilon)O\left(|\nu_j^{n,N}|\right)q_N(\nu_j^{n,N})\right|^2 \leqslant Ks.$$

Since there is nothing random in the second term on the right-hand side of (3.25),

$$\left|\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}\frac{\varepsilon}{\sqrt{\ln(\varepsilon j+A_0)}}D(v^*)\right|^2 \leqslant Ks.$$

By virtue of the independence of W_j^n,

$$E_t^\varepsilon\left|\sqrt{2\varepsilon}\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}\Phi(v^*)W_j^n\right|^2 \leqslant K\varepsilon\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}|\Phi(v^*)|^2EW_j^{n,\tau}W_j^n \leqslant Ks,$$

and similarly,

$$E_t^\varepsilon \left| \sqrt{2\varepsilon} \sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1} [\Phi(v_k^{n,N}) - \Phi(v^*)] q_N(v_k^{n,N}) W_j^n \right|^2$$
$$\leqslant K\varepsilon \sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1} E W_j^{n,\tau} W_j^n \leqslant Ks.$$

Consequently, combining the estimates obtained thus far,

$$E_t^\varepsilon \left| \nu^{\varepsilon,N}(t+s) - \nu^{\varepsilon,N}(t) \right|^2 \leqslant Ks.$$

Therefore, for any $\delta > 0$ (recall that $0 < s < \delta$),

$$\lim_{\delta\to 0} \limsup_{\varepsilon\to 0} E[E_t^\varepsilon \left| \nu^{\varepsilon,N}(t+s) - \nu^{\varepsilon,N}(t) \right|^2] = 0.$$

The tightness of $\{\nu^{\varepsilon,N}(\cdot)\}$ is proved. □

3.3.2. *Weak convergence of* $\nu^{\varepsilon,N}(\cdot)$

We begin with some preliminary calculations. First let us state a lemma, which picks out the effective terms and discards the asymptotically unimportant terms in (3.25).

LEMMA 3.6. *Under the conditions of Theorem 3.5,*

$$\begin{aligned} \nu^{\varepsilon,N}(t+s) - \nu^{\varepsilon,N}(t) = &-\varepsilon \sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1} H v_j^{n,N} \\ &+ \sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1} \frac{\varepsilon}{\sqrt{\ln(\varepsilon j + A_0)}} D(v^*) \\ &+ \sqrt{2\varepsilon} \sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1} \Phi(v^*) W_j^n + o(1), \end{aligned} \tag{3.26}$$

where $o(1) \to 0$ *in probability uniformly in* t *as* $\varepsilon \to 0$.

REMARK 3.7. In view of the lemma, we need only concentrate on the first three terms in (3.26). The rest of the terms in (3.25) are of higher order and can be dropped in the asymptotic analysis.

Proof. We need only show that the last four terms in (3.25) are asymptotically unimportant. First, by the continuity and the boundedness of $f'(\cdot)$, $D(\cdot)$ is bounded and continuous, so

$$\begin{aligned}
&E\left|\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}\frac{\varepsilon}{\sqrt{\ln(\varepsilon j+A_0)}}[D(v_j^{n,N})-D(v^*)]q_N(v_j^{n,N})\right|^2\\
&\leqslant \sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1}\frac{\varepsilon^2}{\sqrt{\ln(\varepsilon j+A_0)}\sqrt{\ln(\varepsilon k+A_0)}}\\
&\times E^{\frac{1}{2}}|[D(v_j^{n,N})-D(v^*)]q_N(v_j^{n,N})|^2E^{\frac{1}{2}}|[D(v_k^{n,N})-D(v^*)]q_N(v_k^{n,N})|^2\\
&\to 0 \text{ as } \varepsilon\to 0.
\end{aligned}$$

By the continuity and the boundedness of $f(\cdot)$, $\Phi(\cdot)$ is bounded and continuous. Owing to the independence of $\{W_j^n\}$,

$$\begin{aligned}
&E\left|\sqrt{2\varepsilon}\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}[\Phi(v_j^{n,N})-\Phi(v^*)]q_N(v_j^{n,N})W_j^n\right|^2\\
&\leqslant 2\varepsilon\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}E|\Phi(v_j^{n,N})-\Phi(v^*)|^2(q_N(v_j^{n,N}))^2E_j^n|W_j^n|^2\\
&\to 0 \text{ as } \varepsilon\to 0.
\end{aligned}$$

Moreover,

$$\begin{aligned}
&E\left|\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}o(\varepsilon)O(|v_j^{n,N}|)q_N(v_j^{n,N})\right|^2\\
&\leqslant \sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1}o(\varepsilon^2)O(E^{\frac{1}{2}}|v_j^{n,N}q_N(v_j^{n,N})|^2E^{\frac{1}{2}}|v_k^{n,N}q_N(v_k^{n,N})|^2)\\
&\to 0 \text{ as } \varepsilon\to 0.
\end{aligned}$$

The lemma is thus proved. □

Next, we state another lemma that gives a functional central limit theorem. It is a variant of the Donsker's invariance theorem.

LEMMA 3.8. *Define*

$$w^\varepsilon(t)=\sqrt{\varepsilon}\sum_{j=0}^{t/\varepsilon-1}W_j^n.$$

Then $w^\varepsilon(\cdot)$ converges weakly to a standard Brownian motion $w(\cdot)$.

Proof. This follows from a standard argument. See Chapter 10.6 of Kushner and Yin (1997). We omit the details here. □

REMARK 3.9. In view of the familiar Slutsky theorem, $\sqrt{2}\Phi(v^*)w^\varepsilon(t)$ converges weakly to a Brownian motion $\sqrt{2}\Phi(v^*)w(\cdot)$ such that the covariance of the limit Brownian motion is $2\Phi(v^*)\Phi^\tau(v^*)t$. It follows from 3.8,

$$\sqrt{2\varepsilon}\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}\Phi(v^*)W_j^n \text{ converges weakly to } \sqrt{2}\Phi(v^*)\int_t^{t+s} dw(u).$$

To proceed, choose a sequence m_ε such that $m_\varepsilon \to \infty$ but $\delta_\varepsilon = \varepsilon m_\varepsilon \to 0$ as $\varepsilon \to 0$. Rewrite the first term on the right-hand side of (3.26) as

$$\begin{aligned}-\varepsilon\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1} H\nu_j^{n,N} &= -\sum_{l\delta_\varepsilon=t}^{t+s}\delta_\varepsilon\frac{1}{m_\varepsilon}\sum_{j=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} H\nu_{lm_\varepsilon}^{n,N}\\ &-\sum_{l\delta_\varepsilon=t}^{t+s}\delta_\varepsilon\frac{1}{m_\varepsilon}\sum_{j=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} H[\nu_j^{n,N}-\nu_{lm_\varepsilon}^{n,N}].\end{aligned} \tag{3.27}$$

Using the techniques in Chapter 8 of Kushner and Yin (1997), it can be shown that the first term on the second line of (3.27) converges to

$$-\int_t^{t+s} H\nu^N(u)du,$$

whereas the second term goes to 0 in probability uniformly in t. Similarly, it can be shown that the term on the second line of (3.26) has the limit

$$\int_t^{t+s}(1/\sqrt{\ln(u+A_0)})D(v^*)du.$$

We summarize this into the following theorem.

THEOREM 3.10. *In addition to the conditions of Theorem 3.5, assume that $-H$ defined in* (3.19) *is a stable matrix. Then $\nu^{\varepsilon,N}(\cdot)$ converges weakly to $\nu^N(\cdot)$ that is the solution of the stochastic differential equation*

$$d\nu^N = \left(-H\nu^N + \frac{1}{\sqrt{\ln(t+A_0)}}D(v^*)\right)dt + \sqrt{2}\Phi(v^*)dw.$$

3.3.3. *The Limit $\nu(\cdot)$*

In the previous section, we have treated $\nu^{\varepsilon,N}(\cdot)$, a truncated process of $\nu^\varepsilon(\cdot)$. In this section, we obtain the convergence of the original process $\nu^\varepsilon(\cdot)$ by letting $N \to \infty$.

The argument here is similar to that of Corollary to Theorem 3.2 of Kushner (1984). Let $P_{\nu(0)}(\cdot)$ and $P^N(\cdot)$ be the measures induced by $\nu(\cdot)$ and $\nu^N(\cdot)$, respectively. Since the differential equation (3.28) is linear, it has a unique solution for each initial condition, so the measure $P_{\nu(0)}(\cdot)$ is unique. For each $T < \infty$, $P_{\nu(0)}(\cdot)$ agrees with $P^N(\cdot)$ on all Borel subsets of the set of paths in $D^r[0, \infty)$ whose values are in S_N for $t \leqslant T$. Note that

$$P_{\nu(0)}\left(\sup_{t\leqslant T}|\nu(t)| \leqslant N\right) \to 1 \text{ as } N \to \infty.$$

This together with the weak convergence of $\nu^{\varepsilon,N}(\cdot)$ to $\nu^N(\cdot)$ yields that $\nu^\varepsilon(\cdot)$ converges weakly to $\nu(\cdot)$. We thus have:

THEOREM 3.11. *Under the conditions of Theorem 3.10, $\nu^\varepsilon(\cdot)$ converges weakly to $\nu(\cdot)$ that is the solution of the stochastic differential equation*

$$d\nu = \left(-H\nu + \frac{1}{\sqrt{\ln(t + A_0)}}D(v^*)\right)dt + \sqrt{2}\Phi(v^*)dw, \tag{3.28}$$

where H is given by (3.19).

REMARK 3.12. Note that the stationary covariance of the diffusion given by (3.28) is

$$\Sigma = 2\int_0^\infty \exp(-Ht)\,\Phi(v^*)\Phi^\tau(v^*)\exp(-H^\tau t)\,dt.$$

The stability of $-H$ implies that the integral above is well defined. Thus loosely $v_k^n - v^*$ is asymptotically normally distributed with mean zero and covariance $(1/\ln(\varepsilon k + A_0))\Sigma$ (see also Chapter 10 of Kushner and Yin, 1997, for related discussion on stochastic approximation algorithms). The scaling factor together with the stationary covariance of the diffusion (3.28) gives us the rate of convergence result.

4. Modifications and extensions

This section considers several extensions of the rates of convergence obtained. First, we look at the case that the energy function is defined on the hypercube. Next, we examine decreasing step-size algorithms with proper scaling. Finally, we consider algorithms involving additional sources of noise.

4.1. ALGORITHMS WITH $\mathcal{E}(\cdot)$ DEFINED ON $[0, 1]^r$

Consider the same algorithm (2.1), but make the modifications that $\mathcal{E}$ is defined on $[0, 1]^r$ and that $\{W_k^n\}$ is a sequence of bounded random variables. Then the convergence can still be obtained, and the rate of convergence remains to be the same as derived in the last theorem.

THEOREM 4.1. *Under the conditions of Theorem 3.11 with the modification that* $\mathcal{E}(\cdot) : [0,1]^r \mapsto \mathbb{R}$, *that* $v^* \in (0,1)^r$, *that* $\{W_k^n\}$ *is bounded with probability 1, that* $f(0) = f(1) = 0$, *and that* $D^r[0,1]$ *replaces* $D^r[0,\infty)$, *then the conclusion of Theorem 3.11 continues to hold.*

Since the proof is similar to that of Theorem 3.11, we omit the details. The assumption $f(0) = f(1) = 0$ guarantees that the limit diffusion $v(\cdot)$ (that is the limit of the interpolation of v_k^n) is stationary; see Wong (1991). We also require v^* in the interior of the hypercube. This is mainly for the rate of convergence analysis. As was proved in Yin et al. (2000), we can ensure that all the iterates belong to $[0,1]^r$. Since $[0,1]^r$ is compact, by virtue of (A1), $f(\cdot)$, $f'(\cdot)$, and $(\partial/\partial v_\alpha)\mathcal{E}(\cdot)$ are all bounded uniformly on $[0,1]^r$. It turns out that for ε sufficiently small, if $v_k^{\iota n} \in [0,1]$, then $v_{k+1}^{\iota n} \in [0,1]$.

REMARK 4.2. In the image estimation problems that we are interested in, we often choose

$$g(u) = \frac{1}{2} + \frac{1}{\pi}\arctan u$$
$$f(v) = \frac{1}{\pi}\cos^2 \pi(v - 0.5)$$

and choose $\mathcal{E}(v)$ to be a quadratic function. Then the conditions posed are all satisfied.

4.2. DECREASING STEP-SIZE ALGORITHMS

First consider a decreasing step-size algorithm of the form

$$\begin{aligned} v_{k+1}^{\iota n} =& v_k^{\iota n} - a_k^{\iota n} F(v_k^{\iota n})\mathcal{E}_v(v_k^{\iota n}) \\ &+ c_k^{\iota n} D(v_k^{\iota n}) + b_k^{\iota n}\Phi(v_k^{\iota n}) W_k^{\iota n}, \end{aligned} \tag{4.1}$$

where

$$\begin{aligned} a_k^{\iota n} &= 1/(\iota n + k)^\gamma, \\ b_k^{\iota n} &= \sqrt{2a_{\iota n+k}}/[\sqrt{\ln((\in + k)^{1-\gamma} - (\iota n)^{1-\gamma} + A_0)}], \\ c_k^{\iota n} &= a_k^{\iota n}/\ln[(\iota n + k)^{1-\gamma} - (\iota n)^{1-\gamma} + A_0], \end{aligned} \tag{4.2}$$

with $A_0 > 1$ and $1/2 < \gamma < 1$. Again, let us fix $\iota = 1$. In this case, we define

$$v_k^n = \sqrt{\ln(k^{1-\gamma} + A_0)}[v_k^n - v^*], \tag{4.3}$$

and

$$
\begin{aligned}
t_k^n &= \sum_{i=1}^{k} \frac{A}{(i+n)^\gamma},\\
m(t) &= \max\{k;\ t_k^n \leqslant t\},\\
\nu^0(t) &= \nu_k^n \text{ for } t \in [t_n, t_{n+1}),\\
\widetilde{\nu}^k(t) &= \nu^0(t+t_n).
\end{aligned}
$$

THEOREM 4.3. *Under the conditions of Theorem 3.11 with the decreasing step sizes given by* (4.2)*, the conclusion of Theorem 3.11 continues to hold with* $\nu^\varepsilon(\cdot)$ *replaced by* $\widetilde{\nu}^k(\cdot)$

REMARK 4.4. We can obtain another decreasing step-size algorithm (4.1) by changing the step sizes given in (4.2) to

$$
\begin{aligned}
a_k^{\iota n} &= 1/(\iota n + k),\\
b_k^{\iota n} &= \sqrt{2a_{\iota n+k}}/[\sqrt{\ln\ln(k+A_0)}],\\
c_k^{\iota n} &= a_k^{\iota n}/\ln\ln(k+A_0).
\end{aligned}
\tag{4.4}
$$

In this case (with $\iota = 1$), define

$$
\begin{aligned}
t_k^n &= \sum_{i=1}^{k} \frac{1}{(i+n)},\\
\nu_k^n &= \sqrt{\ln\ln(k+A_0)}[v_k^n - v^*].
\end{aligned}
\tag{4.5}
$$

Then the desired result still holds.

4.3. NOISY MEASUREMENTS

In applications, additional noise other than the added disturbance $\{W_k^{\iota n}\}$ frequently arises. Such random errors may be due to digitization or noisy observations or a combination of these. Let the additional noise be denoted by $\xi_k^{\iota n}$. Then the algorithm can be modified accordingly as follows:

$$
v_{k+1}^{\iota n} = v_k^{\iota n} - a_k^{\iota n} F(v_k^{\iota n})\mathcal{E}_v(v_k^{\iota n}) + a_k^{\iota n}\xi_k^{\iota n} + c_k^{\iota n} D(v_k^{\iota n}) + b_k^{\iota n}\Phi(v_k^{\iota n})W_k^{\iota n}, \tag{4.6}
$$

where $\{a_k^{\iota n}\}$, $\{b_k^{\iota n}\}$ and $\{c_k^{\iota n}\}$ are either decreasing step-size sequences or constant step-size sequences defined previously. In Yin et al. (2000), we have established the convergence of the algorithm under additional contributing noise. Using the methods developed in this work, we can obtain the rates of convergence of such algorithms. The analysis is similar, and the result is given as follows. Again, for simplicity, we take $\iota = 1$.

THEOREM 4.5. *For the constant-step algorithm with step-size sequence given by* (2.2) (*resp. the decreasing-step algorithm with step size* (4.2)), *in addition to the assumptions of Theorem 3.11* (*resp. Theorem 4.3*), *assume that* $\{\xi_k^n\}$ *is a stationary* φ*-mixing process independent of* $\{W_k^n\}$ *such that*

(a) $E\xi_k^n = 0$, $E|\xi_k^n|^2 < \infty$;

(b) *there exists a sequence of nonnegative real numbers* $\{\rho_n\}$ *such that for each* $k \geqslant j$,

$$E^{1/2}|E_j\xi_k^n - E\xi_k^n|^2 \leqslant \rho_{k-j}, \quad \text{and} \quad \sum_k \rho_k < \infty,$$

where E_m *denotes the conditioning on the* σ*-algebra* $\mathcal{F}_m^n$ *generated by* $\{v_0^n, \xi_i^n, W_i^n;\ i < m\}$.

Then the conclusions of Theorem 3.11 (*resp. Theorem 4.3*) *continue to hold.*

REMARK 4.6. Roughly, the mixing condition requires $\{\xi_j^n,\ j < l\}$ and $\{\xi_j^n,\ j \geqslant l+k\}$ be independent as $k \to \infty$. For definition and discussion of mixing processes, see Ethier and Kurtz (1986, p. 345). It is well known that a stationary φ-mixing process is ergodic. Thus, the condition for the noise $\{\xi_k^n\}$ implies that

$$\frac{1}{k}\sum_{j=m}^{k+m-1} E_m\xi_j^n \to 0 \text{ in probability.}$$

Thus the condition needed for convergence in Yin et al. (2000) is verified. As observed in Yin (1999), this observation noise contributes nothing to the limit stochastic differential equation in the rate of convergence analysis. Thus the most important scaling factor comes from the step of the perturbing noise term not from the noisy observation or measurements.

5. Applications to image processing

In principle, a diffusion network can be used to solve any optimization problem put in the form of minimization of an energy function over $[0, 1]^r$. Since the network performs parallel computations, it is potentially most useful for large-scale problems involving many variables. Examples of such problems arise in image estimation, including *segmentation* (i.e., a partition of an image into a small number of classes) and *restoration* (i.e., recovery of (continuous-valued) image data from corrupted observations). It is noted in Manjunath et al. (1990) that for images using *Markov Random Field* (*MRF*) models, segmentation can be accomplished by minimizing an appropriate function (a *Gibbs distribution* energy function) over the discrete set $\{0, 1\}^r$. In Yin et al. (2000), we used ideas similar to those in Manjunath et al. (1990) to develop diffusion networks (operating over $[0, 1]^r$) for performing image segmentation and restoration. The examples in Yin et al.

(2000) used networks having a fixed step size (without a periodic restart). The fixed step-size results in Section 3 of this paper indicate that a larger step leads to faster convergence (because of the scaling factor in Remark 3.12), but of course, a larger step also leads to a greater approximation error. The results of Section 4.2 indicate that, when following the decreasing step-size schedule of (4.2), there is a slower convergence rate for large k (i.e., when $k^{1-\gamma} < \varepsilon k$), but one would also expect eventually to achieve less approximation error than for fixed step sizes. These tradeoffs suggest that both rapid convergence and small approximation error could be obtained by using initially large but decreasing step sizes combined with a periodic restart. In this section we first compare fixed and decreasing step sizes for segmentation of a noisy two-region image, with test results showing faster convergence to a better result in the case of decreasing step sizes with periodic restart. We then consider a problem of joint segmentation and restoration of a blurred (for a precise definition of blur, see Section 5.2 of this paper, in particular (5.2)) and noisy image, with test results showing slightly better performance for decreasing step sizes than for fixed steps.

5.1. SEGMENTATION OF A NOISY TWO-REGION IMAGES

We first define an image model as in Yin et al. (2000). Let Ω denote the $K \times L$ lattice of pixels on which images are defined, and let the pixels be indexed by $\ell = 1, \ldots, KL$. Suppose that the desired image consists of M constant-intensity regions. Let $\{\mu_m : m = 1, \ldots, M\}$ denote the set of region intensities. We define $X = \{X(\ell) : \ell = 1, \ldots, KL\}$ to be the field of *region labels*. That is, each $X(\ell)$ takes a value in the set $\{1, \ldots, M\}$; and if $X(\ell) = x(\ell)$, then the mean intensity at pixel ℓ is $\mu_{x(\ell)}$. To impose on the model the spatial continuity inherent in image regions, we assume that X has a Gibbs distribution in the form

$$P(X = x) \propto \exp\left(\sum_{c \in C}\{V_c(x(m) : m \in c)\}\right)$$

where c is a subset of $\{1, \ldots, KL\}$ called a *clique*; C is the collection of all cliques; and $V_c(\cdot)$ is some function of x restricted to c. Let $\eta_\ell = \{m : m \text{ is in a clique with } \ell\}$ (η_ℓ is called the *neighborhood* of ℓ). (For example, a commonly-used image model is

$$P(X = x) \propto \exp\left\{-2\beta \sum_{\ell=1}^{KL} \sum_{m \in \eta_\ell} [1 - \delta(x(\ell) - x(m))]\right\}$$

where β is a positive constant, η_ℓ is the set of four or eight pixels nearest to ℓ, and $\delta(\cdot)$ is the Kronecker delta function.)

Suppose that the image is observed in additive white Gaussian noise. That is, the observed image Y is defined by

$$Y(\ell) = \mu_{X(\ell)} + N(\ell), \quad \ell = 1, \ldots, KL,$$

where $N = \{N(\ell)\}$ is a field of i.i.d. Gaussian random variables, each having mean zero and variance σ^2. We assume that X and N are independent. The desired segmentation estimate is the *maximum a posteriori* (*MAP*) estimate of X, given an observed image $Y = y$; that is,

$$\begin{aligned}\widehat{X} &= \arg\max_x\{P(X = x \mid Y = y)\} \\ &= \arg\max_x\{\log(P(Y = y \mid X = x)) + \log(P(X = x))\}.\end{aligned}$$

It is shown in Yin et al. (2000) that the MAP segmentation can be accomplished through use of a diffusion network having energy function

$$\begin{aligned}\mathcal{E}(v) =& \frac{1}{2\sigma^2}\sum_{\ell=1}^{KL}\langle g(\ell), v(\ell)\rangle + \beta\sum_{\ell=1}^{KL}\sum_{m\in\eta_\ell}|v(\ell) - v(m)|^2 \\ &+ \lambda\sum_{\ell=1}^{KL}\{\langle v(\ell), u - v(\ell)\rangle + (1 - \langle v(\ell), u\rangle)^2\}\end{aligned} \tag{5.1}$$

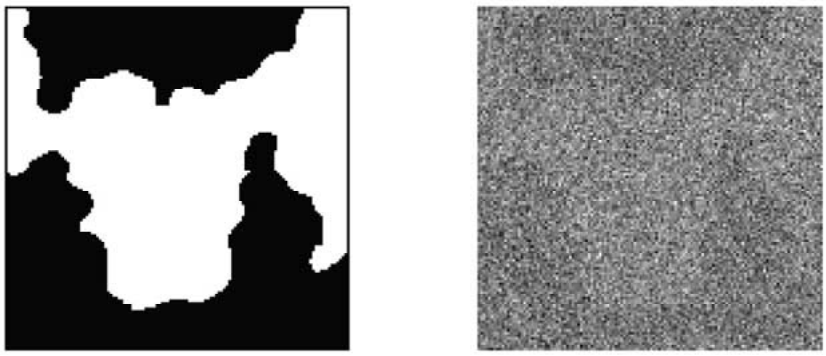
where for each pixel ℓ,

$$\begin{aligned}g(\ell) &= [(y(\ell) - \mu_1)^2, \ldots, (y(\ell) - \mu_M)^2]^\tau \text{ and} \\ v(\ell) &= [v_1(\ell), \ldots, v_M(\ell)]^\tau,\end{aligned}$$

with $v(\ell) = e_m$ (the m^{th} unit vector in $\mathbb{R}^M$) if $x(\ell) = m$; and where λ is a 'large enough' constant that the minimum of (5.1) over $([0, 1]^M)^{KL}$ occurs at a point in the set $\{e_1, \ldots, e_M\}^{KL}$.

A diffusion network having the energy function (5.1) was implemented in MATLAB. In the implementation, the $\{W_{\alpha,k}\}$ sequences in (2.1) were taken to be independent Bernoulli random variables. The function $f(\cdot)$ was set to have the form given in 4.2. In the energy function defined by (5.1), we set $\beta = 0.3$, let η_ℓ consist of the eight nearest neighbors to pixel ℓ, and set $\lambda = 4.5$ (the largest λ value used in the tests in Yin et al. (2000)). For our test image, we set $K = L = 128$, $M = 2$, $\{\mu_1, \mu_2\} = \{1, 2\}$, and $\sigma = 2$. Figure 1 shows the true image x on the left and the observed image y on the right. For this case, a maximum likelihood estimator of region labels would have an error rate of approximately 41%.

For the first test, we used the fixed step size algorithm defined by (2.2) (with ι fixed at 0 and $A_0 = 10$). Using the step size bound given in Yin et al. (2000), we set $\varepsilon = 0.02$. The network was run for 15000 iterations (starting from a randomized $v(\cdot)$), at the end of which each $v(\ell)$ was set to the nearest element of $\{e_1, \ldots, e_M\}$, and $\widehat{X}(\ell)$ was set to j if $v(\ell) = e_j$. The resulting segmentation is shown in Figure 2(a). The pixel error rate in this segmentation is 7.97%. (It might be noted that Dowell (1999) compared the performance of standard sequential implementations of simulated annealing (having in verse linear temperature schedules) and fixed step-size diffusion networks applied to the segmentation of two-region images like



(a) (b)

Figure 1. (a) True image; (b) Noisy observed image.

(a) (b)

Figure 2. (a) Segmentation with fixed step size; (b) Segmentation with decreasing step size and periodic restart.

our test image. In every case, the pixel error rates for the two approaches were similar (with differences generally equal to a fraction of 1%).)

We then tested a decreasing step size algorithm with periodic restart on the same image, using coefficients defined as in (4.2) with $A_0 = 10$, $n = 5000$, and $\gamma = .55$. (To ensure that iterates remained in the hypercube, the a_k^{in} of (4.2) was replaced by $\min\{.02, 1/(\iota n + k)^{\gamma}\}$). Again the network was run for a total of 15000 iterations. The resulting segmentation is shown in Figure 2(b). The pixel error rate in this segmentation is 6.49%.

Figure 3 is a plot of the pixel error rates vs. iteration number for the two cases, showing faster convergence with lower final error for the case of decreasing step size with periodic restart.

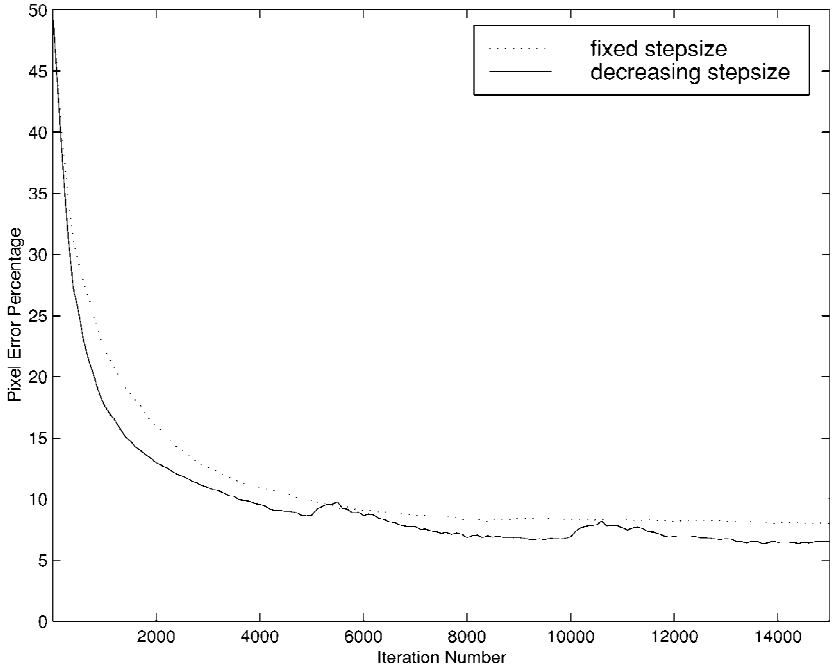


Figure 3. Comparison of pixel error rates.

5.2. RESTORATION AND SEGMENTATION OF A BLURRED AND NOISY IMAGE

In many applications, the observed image is corrupted by both noise and blur. That is, if the true underlying image is G, then the observed image is

$$Y = h * G + N, \tag{5.2}$$

where $h(\cdot)$ is a spatially-invariant blurring filter due, for example, to the imaging system's point-spread function; the '$*$' denotes convolution; and N is a noise term (e.g., see Hokland and Kelly (1996)). It is desired to find an estimate of G, given an observation $Y = y$; i.e., to perform image restoration. A complicating factor is that G is usually nonstationary. For example, in medical ultrasound imaging, G is a diffuse scattering field that can be modelled as independent mean-zero Gaussian random variables having different variances in different tissue regions Hokland and Kelly (1996). As shown in that reference, good estimation results can be obtained for such images by combining restoration and segmentation; that is, by performing segmentation to identify the regions and accounting for the differing variances in different regions during the restoration. In this section, we define and test a diffusion network for restoring and segmenting an image having a model like that used for medical ultrasound images.

Again let X be a two-region field as in Section 5.1. When the region field realization is x, we let the true underlying image be given by $G(\ell) = \sigma_{x(\ell)} N_g(\ell)$ at each pixel ℓ, where $\{\sigma_1^2, \sigma_2^2\}$ are the scatterer variances in the two regions and

N_g is a white Gaussian noise field, independent of X and having mean zero and unit variance at each pixel. (For simplicity, we do not include the specular scattering term used in the model in Hokland and Kelly (1996).) We assume that the observed image is obtained by convolving the true image with a blurring filter and by adding noise N. We further assume that N is also a white Gaussian noise with mean zero and variance σ_N^2 at each pixel, and that N is independent of X and N_g. From this model, it follows that to find the joint MAP estimates of G and X (i.e., joint restoration and segmentation) from an observation y, we need to minimize the energy function

$$U(x, g) = \sum_{\ell=1}^{KL} \{ \frac{1}{2\sigma_N^2} z(\ell)^2 + \ln(\sigma_{x(\ell)}) + \frac{1}{2\sigma_{x(\ell)}^2} g(\ell)^2 + 2\beta \sum_{m \in \eta_\ell} [1 - \delta(x(\ell) - x(m))] \} \tag{5.3}$$

where $z(\ell) = y(\ell) - \sum_m h(\ell - m) g(m)$.

To put this in a form suitable for minimization with a diffusion network, define a vector $v(\ell) = [v_1(\ell), v_2(\ell)]^\tau$ for each pixel ℓ. We use $v_1(\cdot)$ to represent $g(\cdot)$, as follows. Assume that $\sigma_2 > \sigma_1$ and that a $N(0, \sigma^2)$ random variable is effectively confined to the range $\pm 4\sigma$. Then set $v_1(\ell) = \frac{g(\ell) + 4\sigma_2}{8\sigma_2}$. We let $v_2(\cdot)$ govern segmentation by setting $v_2(\ell) = x(\ell) - 1$ (so, each $v_2(\ell)$ is either 0 or 1). Then, with the addition of a constraint term that forces minimization to occur only where each $v_2(\ell)$ is either 0 or 1, the diffusion network energy function is

$$\begin{aligned} \mathcal{E}(v) = \sum_{\ell=1}^{KL} \{ & \frac{1}{\sigma_N^2} [y(\ell) - 4\sigma_2 \sum_m h(\ell - m)(2v_1(m) - 1)]^2 \\ & + [v_2(\ell) \ln(r) + 16(2v_1(\ell) - 1)^2 (r^2 - v_2(\ell)(r^2 - 1))] \\ & + 4\beta \sum_{m \in \eta_\ell} (v_2(\ell) - v_2(m))^2 + \lambda v_2(\ell)[1 - v_2(\ell))] \}. \end{aligned} \tag{5.4}$$

where $r = \sigma_2 / \sigma_1$.

To test the network performance, we used the same two-region image x as in Section 5.1, and set $\sigma_1 = 1$ and $\sigma_2 = 4$. The resulting diffuse scatterer image g is shown in Figure 4(a). This image was blurred with a 2D filter corresponding to a point spread function having value 0.7547 at $(0, 0)$; 0.3396 at $(0, 1)$ and $(1, 0)$; 0.2642 at $(0, -1)$ and $(-1, 0)$; and 0 elsewhere. Gaussian white noise with $\sigma_N = 0.25$ was then added to the blurred image to give the observed image shown in Figure 4(b). The observation was input to a fixed step size diffusion network with energy function (5.4). We again set $\beta = 0.3$, $\lambda = 4.5$, and $A_0 = 10$. The step size bound derived in Yin et al. (2000) depends on the maximum energy function gradient. In this problem, the gradient components corresponding to v_1 have much larger magnitudes than those corresponding to v_2. We found that it was most effective to

use two different step sizes: $\varepsilon_1 = 0.00001$ for components of v_1 and $\varepsilon_2 = 0.005$ for components of v_2. The network was run for 12000 iterations (starting with a randomized $v_2(\cdot)$ and with $v_1(\cdot)$ derived from setting $g(\cdot) = y(\cdot)$). At the end of the run, the restored image was formed by setting $\widehat{G}(\ell) = 4\sigma_2(2v_1(\ell) - 1)$ and the segmented image by setting $\widehat{X}(\ell) = 1 +$ {edge value (0 or 1) nearest to $v_2(\ell)$}. Figure 4(c) shows the restored image $\widehat{G}$. As a measure of restoration efficacy we use *mean-squared error* (*MSE*) *improvement*, defined as $10\log_{10}(\mathrm{MSE}_0/\mathrm{MSE}_f)$ dB, where

$$\mathrm{MSE}_0 = \frac{1}{KL}\sum_{\ell=1}^{KL}[y(\ell) - g(\ell)]^2, \quad \text{and}$$

$$\mathrm{MSE}_f = \frac{1}{KL}\sum_{\ell=1}^{KL}[\widehat{G}(\ell) - g(\ell)]^2. \tag{5.5}$$

For the restoration in Figure 4(c) the MSE improvement is 6.31 dB.

For comparison, we also ran a network with a decreasing step size for the v_2 components, set as $\min\{.005, 1/k^{0.6}\}$ (since the step size for v_1 components must already be very small to keep the iterates in the hypercube, we left it fixed). Figure 4(d) shows the resulting restoration, which has a MSE improvement of 6.45 dB.

Finally, Figure 5(a) shows the segmentation estimate for fixed step size (having a pixel error rate of 2.15%), while Figure 5(b) shows the segmentation estimate with the decreasing v_2 step sizes (having a pixel error rate of 2.09%).

6. Conclusion

This paper has been devoted to rates of convergence of a class of recursive algorithms for global optimization. As demonstrated, the algorithms are useful for many image estimation problems. In Yin (1999), we have studied rates of convergence of Monte-Carlo version of simulated annealing algorithms. In that paper, the rates of convergence is considered for decreasing step-size sequences. The techniques used in the current paper can be adopted to treat simulated annealing algorithms with constant step size. In this paper, the gradient is assumed to be known from digital diffusion network aspects. A Monte Carlo version of the algorithm using a gradient estimator can be constructed and studied. In Lecuyer and Yin (1998), convergence rate results are derived for a stochastic optimization problem where the gradient estimator of the performance measure is available and both the bias and the variance of the estimator depend on the budget devoted to the computation. This idea may be utilized to study the convergence speed of the global optimization algorithm in conjunction with the computational budget. In various applications, one often needs to use projection algorithms or deals with constraints. The resulting recursive algorithms are of constrained type as well. Reference Kushner and Yin (1997) provides extensive discussion on constrained algorithms for stochastic

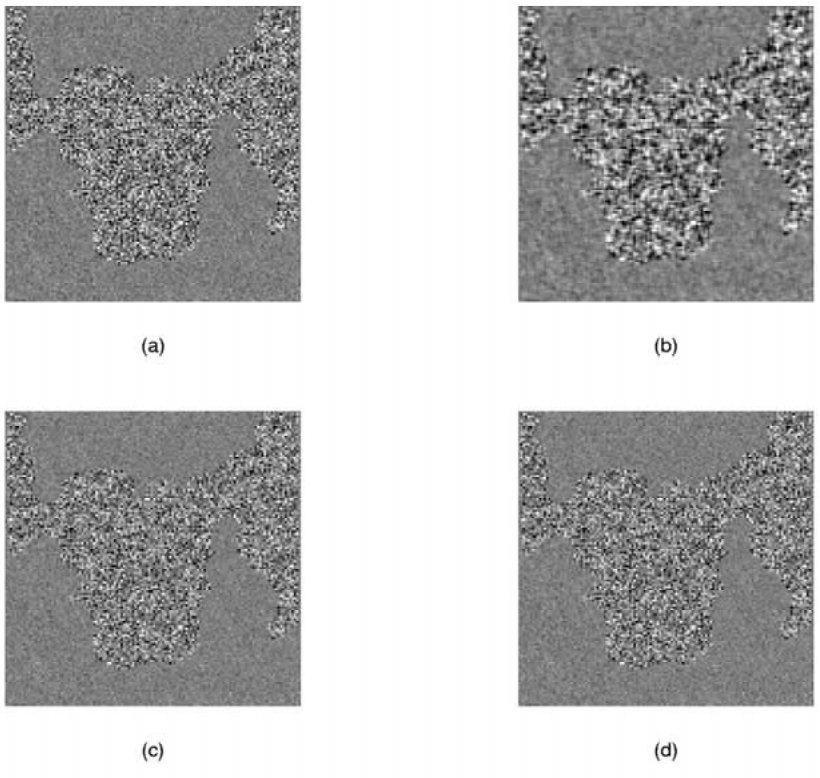


Figure 4. (a) Diffuse scattering field; (b) Noisy and blurred observation; (c) Restoration with fixed step sizes; (d) Restoration with decreasing segmentation step sizes.

Figure 5. (a) Segmentation with fixed step sizes; (b) Segmentation with decreasing segmentation step sizes.

approximation and optimization algorithms with noisy measurements without the perturbing noise W_k^n. The techniques and ideas can be adopted to study the global optimization algorithms that we are considering. The rate of convergence result obtained in this paper is in the spirit of asymptotic normality. Alternatively, one may use large deviations methods to obtain large deviations lower and upper bounds. A challenging problem that is of foremost importance is to improve the convergence rate. One of the ideas points in the direction of using Cauchy-type perturbing noise (without finite moments). However, this requires much more in-depth study and understanding.

Acknowledgements

Research of G. Yin was supported in part by the National Science Foundation under grants DMS-9877090 and DMS-9971608.

References

X. Cai, P. Kelly, and W. B. Gong, Digital diffusion network for image segmentation, *Proc. IEEE Internat. Conf. Image Processing*, 1995.

T. S. Chiang, C.R. Hwang, and S.J. Sheu, Diffusion for global optimization in $\mathbb{R}^n$, *SIAM J. Control Optim.* **25** (1987), 737-752.

M. H. Dowell, *A Parallel Implementation of a Digital Diffusion Network for Image Segmentation*, M.S. Thesis, Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, 1999.

S. N. Ethier and T. G. Kurtz, *Markov Processes, Characterization and Convergence*, Wiley, New York, 1986.

S. B. Gelfand and S. K. Mitter, Recursive stochastic algorithms for global optimization in $\mathbb{R}^d$, *SIAM J. Control Optim.* **29** (1991), 999-1018.

J. Hokland and P. Kelly, Markov models of specular and diffuse scattering in restoration of medical ultrasound images, *IEEE Trans. Ultrasonics, Ferroelectrics and Frequency Control* **43** (1996), 660-669.

G. Kesidis, Analog optimization with Wong's stochastic neural network, *IEEE Trans. Neural Net.* **6** (1995), 258-260.

S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, Optimization by simulated annealing, *Science* **220** (1983), 671-680.

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), 169-185.

H. J. Kushner, *Approximation and Weak Convergence Methods for Random Processes, with applications to Stochastic Systems Theory*, MIT Press, Cambridge, MA, 1984.

H. J. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, New York, 1997.

P. L'Ecuyer and G. Yin, Budget-dependent rate of stochastic approximation, *SIAM J. Optim.* **8** (1998), 217-247.

B. Manjunath, T. Simchony, and R. Chellappa, Stochastic and deterministic networks for texture segmentation, *IEEE Trans. ASSP* **38** 1990.

M. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A.H. Teller, and E. Teller, Equations of state calculations by fast computing machines, *J. Chem. Phys.* **21** (1953), 1087-1091.

E. Wong, Stochastic neural networks, *Algorithmica* **6** (1991), 466-478.
G. Yin, Rates of convergence for a class of global stochastic optimization algorithms, *SIAM J. Optim.*, **10** (1999), 99-120.
G. Yin, P.A. Kelly, and M.H. Dowell, Approximation of an analog diffusion network with applications to image estimation, *J. Optim. Theory Appl.* **107** (2000), 391-414.