SIAM J. CONTROL AND OPTIMIZATION
Vol. 31, No. 4, pp. 1045–1062, July 1993
© 1993 Society for Industrial and Applied Mathematics

STOCHASTIC APPROXIMATION WITH AVERAGING OF THE ITERATES: OPTIMAL ASYMPTOTIC RATE OF CONVERGENCE FOR GENERAL PROCESSES*

HAROLD J. KUSHNER†‡ AND JICHUAN YANG†§

Abstract. Consider the stochastic approximation algorithm

$$X_{n+1} = X_n + a_n g(X_n, \xi_n).$$

In an important paper, Polyak and Juditsky [*SIAM J. Control Optim.*, 30 (1992), pp. 838–855] showed that (loosely speaking) if the coefficients a_n go to zero slower than $O(1/n)$, then the averaged sequence $\sum_{i=1}^{n} X_i/n$ converged to its limit, at an optimum rate, for any coefficient sequence. The conditions were rather special, and direct constructions were used. Here a rather simple proof is given that results of this type are generic to stochastic approximation, and essentially hold any time that the classical asymptotic normality of the normalized and centered iterates holds. Considerable intuitive insight is provided into the procedure. Simulations have well borne out the importance of the method.

Key words. stochastic approximation, stochastic approximation with averaged iterates, rate of convergence for stochastic algorithms

AMS subject classifications. 62L20, 60F05, 62E25

1. Introduction.

Consider the stochastic approximation (SA)

$$X_{n+1} = X_n + a_n g(X_n, \xi_n), \tag{1.1}$$

where $0 < a_n \to 0, \sum_n a_n = \infty$ and $\{\xi_n\}$ is some "noise" sequence. Suppose that for some $\theta, X_n \to \theta$ either with probability one or weakly. Then, under appropriate conditions $(X_n - \theta)/\sqrt{a_n}$ converges in distribution to a normal random variable with mean zero and some covariance matrix V_0. The matrix V_0 is often considered to be a measure of the "rate of convergence," taken together with the scale factors or gains $\{a_n\}$.

Suppose that $a_n \to 0$ "slower" than $1/n$. In particular, suppose that

$$a_n/a_{n+1} = 1 + o(a_n). \tag{1.2}$$

Define

$$\overline{X}_n = \frac{1}{n}\sum_{1}^{n} X_i. \tag{1.3}$$

Then, in a sequence of fundamental papers, Polyak [1] and Polyak and Juditsky [2] showed that $\sqrt{n}(\overline{X}_n - \theta)$ converged in distribution to a normal random variable with mean zero and covariance V, where V was the smallest possible in an appropriate sense. V did not depend on $\{a_n\}$, provided that (1.2) held. This relaxation of conditions on $\{a_n\}$ allows it to be "relatively large" and is quite important in applications.

* Received by the editors July 1, 1991; accepted for publication (in revised form) February 28, 1992.

† Division of Applied Mathematics, Brown University, Box F, Providence, Rhode Island 02912.

‡ The work of this author was supported by Air Force Office of Scientific Research grant AFOSR 89-0015, and Army Research Office grant ARO-DAAL 03-86K-0171.

§ The work of this author was partially supported by National Science Foundation grant ECS-8913351.

Downloaded 02/23/22 to 222.29.96.30 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 02/23/22 to 222.29.96.30 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Simulations have supported the theoretical conclusions and have shown the clear superiority of the use of averages of the type (1.3) over X_n directly. This superiority would not be true if a_n decreased as $O(1/n)$, and in this latter case, it was known for a long time that the asymptotic rates are the same for $\{\overline{X}_n\}$ and $\{X_n\}$. In the past, a great deal of attention was given to the problem of choosing optimal sequences $\{a_n\}$ via "adaptive" and generally unreliable means. The importance of this problem is now much reduced. The results in [1], [2] are similar in spirit to the approach of Ruppert [3] for a one-dimensional case.

The proofs in [1], [2] were by direct construction. They involved detailed expansions and estimates and made no use of prior results in SA. The function $g(\cdot)$ needed to be "smooth" and the conditions on the noise were restrictive, being essentially that $\{\xi_n\}$ were either i.i.d. or martingale differences, depending on the case. The conditions on the noise were weakened by Yin [4], [5] who allowed certain "mixing sequences," but the proofs were still complicated and made no use of prior results in SA.

In this paper, it will be seen that a very simple use of prior results in SA allows us to get results of the above type under condition of considerable generality. In addition, the approach sheds more light on the reasons why averaged estimators such as (1.3) work well. In general, we use sums of the type (1.3), but where the lower index of summation goes to infinity as $n \to \infty$. We will work with several such averages.

In §2, we show that a very useful averaging result can be readily obtained via a weak convergence method under quite broad conditions. This result uses a "minimal window" of iterates, fewer than in (1.3), yet is quite useful in practice. Insight into the reasons why the averaging method works is obtained via an examination of a "two time scale SA" representation of (1.1), (1.3) in a simple case. The "window of averaging" is extended in §3.

2. The basic convergence theorem for the averaged iterates. Define the "interpolated time" $t_n = \sum_0^{n-1} a_i$, with $t_0 = 0$, and its "inverse" $m(t) = \max\{n : t_n \le t\}$. For notational simplicity and without loss of generality, set $\theta = 0$. For each $n \ge 0$, define the interpolated processes $X^n(\cdot)$ and $U^n(\cdot)$ by

$$\left.\begin{aligned} X^n(t) &= X_{n+i} \\ U^n(t) &= X_{n+i}/\sqrt{a_{n+i}} \end{aligned}\right\} \text{ for } t \in [t_{n+i} - t_n, t_{n+i+1} - t_n), i \ge 0.$$

Let $\Rightarrow$ denote weak convergence in the Skorohod topology on $D^r[0,\infty)$ [6], [7]. In Theorem 2.1, we will use the following assumption.

Assumption A2.1. There is a matrix G whose eigenvalues lie in the open left half plane and a positive definite symmetric matrix R_0 such that $X^n(\cdot) \Rightarrow$ zero process and $U^n(\cdot) \Rightarrow U(\cdot)$, where $U(\cdot)$ is the stationary solution to

$$dU = GU\,dt + R_0^{1/2}\,dw. \tag{2.1}$$

Comment on (A2.1). We prefer to state the condition in the form of (A2.1) since so many different sets of conditions imply (A2.1). Also, the main aim here is to show that a standard weak convergence result can be used to get an optimal rate of convergence under any of the sets of conditions which guarantee the usual limit result (A2.1). The references [8], [9], [11]–[13] contain various sets of conditions which guarantee (A2.1).

Downloaded 02/23/22 to 222.29.96.30 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

For $t > 0$, define $Z^n(\cdot)$ by

$$Z^n(t) = \frac{1}{\sqrt{t/a_n}} \sum_{i=n}^{n+t/a_n} X_i. \tag{2.2}$$

In sums of type $\sum_{\alpha}^{\beta}$ for real α, β, we always use the integer parts of α, β.

A basic convergence theorem.

THEOREM 2.1. *Assume* (1.2) *and* (A2.1) *and define* $V = G^{-1}R_0(G')^{-1}$. *For each* t, $Z^n(t)$ *converges in distribution to a random variable with mean zero and covariance* $V_t = V + O(1/t)$.

Proof. Define the processes

$$\tilde{Z}^n(t) = \frac{1}{\sqrt{t}} \int_0^t U^n(s)ds, \qquad \tilde{Z}(t) = \frac{1}{\sqrt{t}} \int_0^t U(s)ds.$$

By the weak convergence in (A2.1), $\tilde{Z}^n(\cdot) \Rightarrow \tilde{Z}(\cdot)$. Define the covariance matrix $R(s) = EU(t)U'(t+s)$, where $U(\cdot)$ is the stationary solution to (2.1). Since $R(s) \to 0$ exponentially as $s \to \infty$,

$$\begin{aligned} \text{cov } \tilde{Z}(t) &= \frac{1}{t} \int_0^t \int_0^t R(s-\tau)dsd\tau \\ &= \int_{-\infty}^{\infty} R(s)ds + O(1/t), \end{aligned}$$

but $\int_{-\infty}^{\infty} R(s)ds = G^{-1}R_0(G^{-1})'$.

The basic result on the character of the averaged iterates is obtained by relating $Z^n(t)$ to $\tilde{Z}^n(t)$.

Write

$$\frac{a_n}{a_{n+i}} = 1 + \delta_{n,i}. \tag{2.3}$$

Then (1.2) implies that for any $t < \infty$,

$$\max\{i - n : 0 \le t_i - t_n \le t\} \cdot a_n/t \xrightarrow{n} 1. \tag{2.4}$$

Equation (2.4) will be heavily used. Note that (2.4) would not hold if $a_n = O(1/n)$. Equation (2.4) follows from (1.2) as follows: Let $i > n$. Then $a_i/a_n = \prod_n^i(1 + o(a_j))$. If $\sum_n^i a_j \le t$, then the ratio goes to unity, uniformly in such i, as $n \to \infty$.

Using the "piecewise constant" definition of $U^n(\cdot)$, we have (modulo "end terms") for $i \ge n$

$$\begin{aligned} \sqrt{t}\tilde{Z}^n(t) &= \sum_{i:t_i - t_n \le t} (X_i a_i^{-1/2})a_i \\ &= \sum_{i:t_i - t_n \le t} X_i(a_i^{1/2} - a_n^{1/2}) + a_n^{1/2} \sum_{i:t_i - t_n \le t} X_i. \end{aligned}$$

(Alternatively, the sums can be written as $\sum_{m(t_n)}^{m(t_n+t)}$.) By the weak convergence of $U^n(\cdot)$ in (A2.1) and (2.4), the first sum on the right goes to zero in probability as

$n \to \infty$. By the same weak convergence and (2.4), the second sum on the right is asymptotically equivalent (in distribution) to

$$a_n^{1/2} \sum_{i=n}^{n+t/a_n} X_i. \tag{2.5}$$

This and the weak convergence $\tilde{Z}^n(t) \Rightarrow \tilde{Z}(t)$ yield the theorem. □

Discussion of the theorem.

(a) On the optimality of the "rate of convergence." Suppose that $U_n = X_n/\sqrt{a_n}$ converged in distribution to a normally distributed random variable $\hat{U}$ with mean zero. It is common to consider the covariance of $\sqrt{a_n}\hat{U}$ as a "measure of the rate of convergence" or "asymptotic errors." Then the best value of a_n is $O(1/n)$. Suppose that $a_n = A/n$, for A a positive definite matrix. Then, under appropriate conditions (see, e.g., [8], [9]), $U^n(\cdot) \Rightarrow \tilde{U}(\cdot)$, where $\tilde{U}(\cdot)$ is the stationary solution to

$$d\tilde{U} = \left(\frac{I}{2} + AG\right)\tilde{U}dt + AR_0^{1/2}dw, \tag{2.6}$$

where G and R_0 are as in (A2.1), and it is supposed that $(I/2 + AG)$ is a stable matrix. If we optimize the trace of the covariance matrix of (2.6) over A, we get the best value of A as

$$A = -G^{-1}.$$

With this value of A, the covariance of $\tilde{U}(0)$ is just the V used in Theorem 2.1. In this sense, the result in [1], [2] and of Theorem 2.1 is optimal. Let us note that the fact that $U^n(\cdot) \Rightarrow U(\cdot)$ satisfying (2.1), rather than (2.6), is of crucial importance. The integral of the correlation function of (2.6) depends on A. Again we emphasize that Theorem 2.1 requires that $a_n \to 0$ slower than $O(1/n)$. The result of Theorem 2.1 holds for some very complicated SAs, e.g., ones which arise due to distributed and asynchronous processing [13].

(b) The "window" of averaging. The value of t can be made as large as desired in (2.2), and can go to infinity slowly with n. More will be said about this in the next section. A two-sided average

$$\sqrt{\frac{a_n}{t_1 + t_2}} \sum_{n-t_1/a_n}^{n+t_2/a_n} X_i$$

can be used in lieu of (2.2) with the same results. For this case, the proof is nearly identical to the one given.

In (2.2), the "window" of the averaging is $O(1/a_n)$ as opposed to $O(n)$ in (1.3). Theorem 2.1 implies that the order $O(1/a_n)$ is the *smallest* which can be used. Suppose that $a_n = 1/n^\gamma, \gamma \in (0,1)$. Then as $\gamma \to 0$, the minimal window of averaging decreases. Loosely speaking, for smaller rates of decrease of $\{a_n\}$, there is more "oscillation" of the iterates $\{X_n\}$ about the limit point, and less averaging is needed. This point will be supported by the following discussion of the singularly perturbed SA.

Downloaded 02/23/22 to 222.29.96.30 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 02/23/22 to 222.29.96.30 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

The theorem shows that the improvement, due to averaging, is a *natural property* of SA and is essentially a consequence of weak convergence of the normalized process $U^n(\cdot)$.

(c) Relationships between the use of (2.2) and a two time scale SA. For additional motivation and insight concerning the averaging method, let us rewrite the iteration for (X_n, U_n) as a two time scale or "singularly perturbed" SA. The following discussion is purely heuristic. Hence, for simplicity of presentation, we use a linear and one-dimensional model. For $A > 0$ and $\gamma \in (0,1)$, define $\{X_n\}$ by

$$X_{n+1} = \left(1 - \frac{AG}{n^\gamma}\right) X_n + \frac{A\xi_n}{n^\gamma}.$$

Define $\overline{U}_n = \sum_1^n X_i/\sqrt{n}$. Then, putting the iterations for $\overline{U}_n$ and X_n on the same time scale, we can write

$$\frac{1}{n^{1-\gamma}}(X_{n+1} - X_n) = -\frac{AGX_n}{n} + \frac{A\xi_n}{n}, \tag{2.7a}$$

$$\overline{U}_{n+1} - \overline{U}_n = -\frac{\overline{U}_n}{2n}\left(1 + O\left(\frac{1}{n}\right)\right) + \frac{X_{n+1}}{\sqrt{n+1}}. \tag{2.7b}$$

$\overline{U}_n$ is just $\sqrt{n}$ times the averaged value $\sum_1^n X_i/n$, and a quantity whose asymptotic variance is of interest. Equation (2.7) can be viewed as a two time scale SA.

We can make a similar heuristic argument if $\overline{U}_n$ is replaced by $\sqrt{t}Z^n(t)$. Define a new interpolation time $\tilde{t}_n = \sum_{i=1}^n 1/i$. Define the new continuous time interpolation $\tilde{X}^n(\cdot)$ by $\tilde{X}^n(t) = X_{n+i}$ on $[\tilde{t}_{n+i} - \tilde{t}_n, \tilde{t}_{n+i+1} - \tilde{t}_n), i \geq 0$. In this new time scale, $\{X_n\}$ is "squeezed" or compressed more than it was in the $\{t_n\}$ scale and $\tilde{X}^n(\cdot)$ has a smaller correlation than $X^n(\cdot)$. This "smaller correlation" suggests that an averaging method will yield an improved result. Note that this two time scale effect doesn't hold if $a_n = O(1/n)$.

A two time scale continuous parameter system that is loosely analogous to (2.7) is

$$\varepsilon dz^\varepsilon = A_{11}z^\varepsilon dt + dw_1 \tag{2.8}$$

$$dx^\varepsilon = A_{22}x^\varepsilon dt + A_{12}z^\varepsilon dt + dw_2,$$

where ε is small. Under suitable stability conditions, [10, Chap. 10], $\int_0^t z^\varepsilon(s)ds$ converges weakly to a Wiener process. Hence, we might expect that, with (2.7), the function of t defined by

$$\frac{1}{\sqrt{n}}\sum_{i=0}^{nt} X_i$$

might also converge weakly to a Wiener process with covariance matrix V. Indeed, a similar result is proved in the next section.

Constant gain coefficients. Replace (1.1) by

$$X_{n+1}^\varepsilon = X_n^\varepsilon + \varepsilon g(X_n^\varepsilon, \xi_n), \qquad \varepsilon > 0. \tag{2.9}$$

Define $X^\varepsilon(\cdot)$ and $U^\varepsilon(\cdot)$ by $X^\varepsilon(t) = X_n^\varepsilon, U^\varepsilon(t) = X_n^\varepsilon/\sqrt{\varepsilon}$ on $[n\varepsilon, n\varepsilon+\varepsilon)$. Suppose that there are $t_\varepsilon \xrightarrow{\varepsilon} \infty$ such that $U^\varepsilon(t_\varepsilon + \cdot) \Rightarrow U(\cdot)$, a stationary process which satisfies (2.1). Such results are proved in [11]. Fix $t > 0$ and let $t_i \geq 0$ such that $t = t_1 + t_2$. Define

$$Z^\varepsilon(t) = \sqrt{\frac{\varepsilon}{t}} \sum_{(t_\varepsilon - t_2)/\varepsilon}^{(t_\varepsilon + t_1)/\varepsilon} X_i.$$

Then, following the argument of Theorem 2.1 yields that for each t, $Z^\varepsilon(t)$ converges in distribution to a normally distributed random variable with mean zero and covariance $V + O(1/t)$.

A note on computation. In applications, averages of the type

$$\frac{1}{m_2(n) - m_1(n)} \sum_{m_1(n)}^{m_2(n)-1} X_i$$

are usually preferred to (1.3), where $m_i \to \infty$ and $m_2(n) - m_1(n) \to \infty$ as $n \to \infty$. Such averages cannot, in general, be computed recursively. However, memory requirements can be reduced by appropriate grouping of the iterates and selection of the times of updating.

3. Increasing the window of averaging in (2.2). In this section, we will see how fast we can let $t \to \infty$ and prove the above assertion concerning convergence to Wiener process.

For a sequence $n \geq q_n \to \infty$, define $M^n(t)$ by

$$M^n(t) = \frac{1}{\sqrt{q_n}} \sum_{i=n}^{n+q_n t} X_i. \tag{3.1}$$

We could use $\sum_{i=n-q_n' t}^{n+q_n'' t}$ in (3.1) where $q_n'' + q_n' = q_n$ and $q_n > \epsilon n$ for some $\epsilon > 0$, with the same end results. This is, in fact, a practical case since we would often wish to delete some initial fraction of the iterates from the averaging. To relate this last form to (1.3), let $t = 1$, $q_n = n$, $q_n' = (1-\alpha)n, \alpha < 1$. Then we get

$$\frac{1}{n} \sum_{i=\alpha n}^{(1+\alpha)n} X_i.$$

Theorem 2.1 dealt with the case $q_n = O(1/a_n)$. In this section we need to assume that

$$\text{(A3.1)} \qquad q_n a_n^{3/2} \to 0, \quad q_n a_n \to \infty, \quad q_n \leq k_0 n \quad \text{for some } k_0 < \infty.$$

Thus if $q_n = n$ and $a_n = 1/n^\gamma$, then $\gamma \in (2/3, 1)$ is needed. The result in [2], [3] required only $\gamma \in (1/2, 1)$, but our conditions on the noise and dynamics are more general. Under stronger conditions, the method of the following theorem can be carried through with $\gamma \in (1/2, 1)$. In the next section, it will be shown that $M^n(\cdot) \Rightarrow w(\cdot)$, a Wiener process with covariance matrix Vt, thus supporting the assertion at the end of the last section. In order to extend the "window of averaging" beyond t/a_n, we need $q_n a_n \to \infty$.

Downloaded 02/23/22 to 222.29.96.30 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 02/23/22 to 222.29.96.30 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Discussion of the noise processes. Two types of noise processes are considered. For the first, the sequence $\{\xi_n\}$ is "exogenous." Loosely speaking, the evolution of $\{X_n\}$ does not affect $\{\xi_n\}$. For the second, or "state dependent" noise, (X_n, ξ_n) is jointly Markov. There is a transition function $p(\xi, \cdot|x)$ such that $P\{\xi_{n+1} \in A|\xi_n = \xi, X_n = x\} = p(\xi, A|x)$. For each n and x, define the Markov process $\{\xi_j(x), j \geq n\}$ with initial condition $\xi_n(x) = \xi_n$ and transition function $p(\xi, \cdot|x)$. Such models were introduced in [11], [14], and also used in [9], [12].

The following additional conditions will be used. Let E_n denote the expectation conditioned on $\{X_i, i \leq n, \xi_i, i < n\}$.

Condition A3.2. There is a continuously differentiable "centering" function $\overline{g}(\cdot)$ such that with the definition $\psi_j(x) = g(x, \xi_j) - \overline{g}(x)$ (for the exogenous noise) and $\psi_j(x) = g(x, \xi_j(x)) - \overline{g}(x), j \geq n$ (for the state dependent noise) we have for each n and x,

$$\sum_{j=n}^{\infty} a_j E_n \psi_j(x) = O(a_n) \tag{3.2}$$

where $O(a_n)$ is uniform in n, ω, x (where ω is the canonical point of the sample space), and $\psi_j(x)$ is bounded. In (3.2), for the state dependent noise case, the initial condition of $\{\xi_j(x), j \geq n\}$ is $\xi_n(x) = \xi_n$, following the usage in [12], [14].

Condition A3.3. $\overline{g}(x) = Gx + \delta g(x)$, where G has its eigenvalues in the open left half plane and $|\delta g(x)| = O(|x|^2)$.

Condition A3.4. $\sum_{j=n}^{\infty} a_j E_n[\psi_j(x) - \psi_j(y)] = O(a_n)|y - x|$, $|g(x, \xi)| \leq O(1)[|x| + 1]$.

Comments on the conditions (A3.2), (A3.4). Such conditions were initially introduced in [11], [15] and have been used in many of the other references; e.g., [4], [5], [9], [11], [12], [13]. They are essentially conditions on the "mixing rate" of the processes. A simple example of (3.2) is where the noise is "exogenous," $\overline{g}(x) = Eg(x, \xi)$ is smooth, $\{\psi_j(x)\}$ is bounded and $\{\xi_n\}$ satisfies a mixing condition with a sufficiently fast mixing rate. Many specific examples are shown in [9], [12]. In [9], the sum in (A3.2) is the solution to the Poisson equation, and then (A3.4) is a condition on the smoothness of that solution. For the state dependent noise case, the transition kernel $p(\xi, \cdot|x)$ often assures that $\int g(x, \xi')p(\xi, d\xi'|x)$ is smooth enough so that (A3.4) holds [11].

We could replace $O(a_n)$ in (A3.2) by $O(a_n)\hat{g}(x)$ where $\hat{g}(x)$ has an appropriate growth rate, but we prefer to keep the development simple.

Stability of (1.1). A main problem in extending the window of averaging in Theorem 2.1 concerns the tightness of $\{X_n/\sqrt{a_n}\}$. Such results are basic to the proofs of (A2.1). In fact, given such tightness, straightforward averaging methods can often be used to get (A2.1). Theorem 3.1 requires the following bounds in Lemma 3.1, and to get that the following stability condition will be used. Recall that we use the assumption that the limit point is $\theta = 0$ for notational convenience and without loss of generality.

Condition A3.5. There is a nonnegative continuous function $V(\cdot)$ whose first and second mixed partial derivatives exist and are continuous. For some positive definite symmetric matrix A and some $\gamma > 0, K < \infty$,

$$V(x) = x'Ax + o(|x|^2),$$

$$V_x'(x)\overline{g}(x) \leq -\gamma V(x), \qquad |V_x(x)|^2 \leq KV(x),$$

$$|g(x,\xi)|^2 \le K(V(x)+1).$$

and $V_{xx}(x)$ is uniformly bounded.

Let $\mathcal{F}_m$ denote the minimal σ-algebra which measures $\{X_i, i \le m; \xi_i, i < m\}$.

LEMMA 3.1. *Assume* (A3.2)–(A3.5). *Then* $\{E|X_n|^4/a_n^2\}$ *is bounded. Let* $k < \infty$. *For any* $\mathcal{F}_m$*-stopping time* q *with values in* $[n, n+kn]$, *we have* $E|X_q|^2/a_n = O(1)$, *uniformly in* n *and in* q *in the given class. The bounds also hold for the* Y_i^n *defined by* (3.7).

Proof. The proof can be seen in the Appendix.

The following theorem gives us the largest window of averaging. It is largest in the sense that if $q_n = O(n)$, then the window is $O(n)$. The mutual independence of the increments of the Wiener process, which is the limit in the theorem, sheds additional light on the time scales which are used.

THEOREM 3.1. *Under* (1.2), (A2.1) *and* (A3.1)–(A3.5), $M^n(\cdot) \Rightarrow W(\cdot)$, *a Wiener process with covariance* Vt.

The proof will be divided into several parts. First we will show that it is sufficient to replace (1.1) by a simpler iteration. Then tightness of $\{M^n(\cdot)\}$ in the Skorohod topology is shown, and finally we prove that the limit of any weakly convergent subsequence is the asserted Wiener process.

Proof. It is notationally easier to do the proof if the rate at which $a_n \to 0$ is very slow. We will work with the additional assumption (see (1.2), (2.4))

$$(*) \qquad \sup_{0 \le i \le kn} |\delta_{n,i}| \xrightarrow{n} 0$$

for each $k < \infty$. This will allow us to replace a_i by a_n when $0 \le i - n \le kn$. The modifications needed for the general case will be stated in Part 7.

Part 1. (In this part, (A3.1) can be replaced by $q_n a_n^2 \to 0$.) Define the quantity

$$\Pi(n,j) = \prod_{i=n}^{j}(I + a_i G), j \ge n, \qquad \Pi(n,n-1) = I.$$

Write (1.1) in the form

$$\begin{aligned} X_{n+1} &= X_n + a_n\overline{g}(X_n) + a_n\psi_n(X_n) \\ &= X_n + a_n G X_n + a_n \delta g(X_n) + a_n\psi_n(X_n). \end{aligned} \tag{3.3}$$

Then

$$\begin{aligned} X_{n+m+1} &= \Pi(n,n+m)X_n + \sum_{j=n}^{n+m}\Pi(j+1,n+m)a_j\delta g(X_j) \\ &\quad + \sum_{j=n}^{n+m}\Pi(j+1,n+m)a_j\psi_j(X_j) \\ &\equiv Q^1_{n,n+m} + Q^2_{n,n+m} + Y^n_{m+1} \end{aligned} \tag{3.4}$$

where the $Q^i_{n,j}$ and Y^n_m are defined in the obvious way.

It will be shown first that

$$\overline{Q}^i_n = \frac{1}{\sqrt{q_n}}\sum_{j=n}^{n+q_n t} E|Q^i_{n,j}| \xrightarrow{n} 0, \qquad i = 1,2. \tag{3.5}$$

Downloaded 02/23/22 to 222.29.96.30 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

This will allow us to drop the Q^1 and Q^2 terms in (3.4). The stability of G implies that there are $\lambda > 0$ and $c < \infty$ such that

$$\|\Pi(n,j)\| \le ce^{-\lambda(t_j - t_n)}, \qquad j \ge n. \tag{3.6}$$

Thus

$$E|Q^1_{n,j}| \le ce^{-\lambda(t_j - t_n)} E|X_n|.$$

Using this bound and the estimate $E|X_n| = O(a_n^{1/2})$ from Lemma 3.1 yields

$$\overline{Q}^1_n \le O(1) \sum_{j=n}^{n+q_n t} e^{-\lambda(t_j - t_n)} O(\sqrt{a_n})/\sqrt{q_n}.$$

By (A3.1), $1/\sqrt{a_n q_n} \to 0$. Using this and (2.3), (2.4) yields that

$$\overline{Q}^1_n = O(1) \int_0^\infty e^{-\lambda s} ds/\sqrt{q_n a_n} \to 0.$$

Next we evaluate $Q^2_{n,n+m}$. For $m \in [n, kn]$, by Lemma 3.1 we have $E|\delta g(X_m)| = O(a_n)$. This and (3.6) yield

$$\begin{aligned} E|Q^2_{n,n+m}| &\le O(1) \sum_{j=n}^{n+m} e^{-\lambda(t_{n+m} - t_{j+1})} a_j E|\delta g(X_j)| \\ &= O(1) a_n \int_0^\infty e^{-\lambda s} ds, \end{aligned}$$

which yields $E|Q^2_{n,n+m}| = O(a_n)$, $m \le kn$. This implies that $\overline{Q}^2_n \le \sqrt{q_n t} O(a_n)$, which goes to zero as $n \to \infty$ by (A3.1).

Thus, to prove the theorem we can replace $\{X_m, m \ge n\}$ by the $\{Y^n_m, m \ge n\}$ process, which can be defined by

$$Y^n_{m+1} = (I + a_m G) Y^n_m + a_m \psi_m(X_m), m \ge n, \tag{3.7}$$

where we define $Y^n_n = 0$. Note that the stability of G in (A3.3) and the boundedness of $\{\psi_m(X_m)\}$ imply that $(Y^n_m, m \ge n)$ is bounded uniformly in n.

Part 2. Let $k > 0$ and let q be an $\mathcal{F}_m$-stopping time with values in $[n, n + kn]$. We next prove that

$$E|Y^n_q (E_q Y^n_j)'| = O(a_n^{3/2}) + O(a_n) E e^{-\lambda(t_j - t_q)}, \tag{3.8}$$

for $kn \ge j \ge q$, where j and q are integers. A perturbed test function method will be used. For $n + kn \ge j \ge n$, define the "perturbations"

$$\delta Y^n_j = \sum_{i=j}^\infty a_i E_j \psi_i(X_j) = O(a_j) \tag{3.9}$$

$$\tilde{Y}^n_j = Y^n_j + \delta Y^n_j$$

Downloaded 02/23/22 to 222.29.96.30 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

where the $O(a_j)$ value is due to (3.2). Note that the argument of the $\psi_i(\cdot)$ in (3.9) is X_j, the state at the lower index of summation.

Note the following:

$$E_j Y_{j+1}^n - Y_j^n = a_j G Y_j^n + a_j E_j \psi_j(X_j), \tag{3.10a}$$

$$\begin{aligned} E_j \delta Y_{j+1}^n - \delta Y_j^n &= -a_j E_j \psi_j(X_j) \\ &\quad + \sum_{i=j+1}^{\infty} a_i E_j [\psi_i(X_{j+1}) - \psi_i(X_j)] \\ &= -a_j E_j \psi_j(X_j) + S_j, \end{aligned} \tag{3.10b}$$

where S_j is defined in the obvious way. By (A3.4),

$$|S_j| = O(a_j^2)[|X_j| + 1]. \tag{3.11}$$

Combining (3.10a), (3.10b) yields

$$E_j \tilde{Y}_{j+1}^n = (I + a_j G)\tilde{Y}_j^n + S_j - a_j G \delta Y_j^n. \tag{3.12}$$

Solving (3.12) yields

$$E_q \tilde{Y}_j^n = \Pi(q, j-1)\tilde{Y}_q^n + \sum_{i=q}^{j-1} \Pi(i, j-1)[E_q S_i - a_i E_q G \delta Y_i^n].$$

Hence, with the estimate $\delta Y_j^n = O(a_j)$ and (3.11), we can write

$$\begin{aligned} |E_q Y_j^n| &= O(1) e^{-\lambda(t_j - t_q)} |Y_q^n| \\ &\quad + O(1) \sum_{i=q}^{j-1} \Pi(i, j-1) a_i^2 (E_q |X_i| + 1) + O(a_n) \\ &= O(1) e^{-\lambda(t_j - t_q)} |Y_q^n| + O(a_n). \end{aligned} \tag{3.13}$$

By Lemma 3.1,

$$E|Y_q^n|^2 = O(a_n). \tag{3.14}$$

Now, combining (3.13) and (3.14) yields (3.8). Equation (3.13) will be used frequently in the sequel.

Part 3. Define

$$F^n(t) = \frac{1}{\sqrt{q_n}} \sum_{i=n}^{n+q_n t} Y_i^n.$$

By the results in Part 1, it is sufficient to prove the theorem for $F^n(\cdot)$ replacing $M^n(\cdot)$.

Tightness of $\{F^n(\cdot)\}$. Let $k < \infty$. Let $r(n)$ be a $\mathcal{F}_m$-stopping time, with values in $[n, n + kq_n]$. To prove tightness, it is sufficient ([7, Thm. 8.6] or, equivalently, [12, Thm. 3.3]) if $\sup_n E|F^n(t)| < \infty$ for each $t > 0$ and

$$\lim_{\delta \to 0} \overline{\lim_n} \sup_{r(n)} E|F^n(t_{r(n)} + \delta - t_n) - F^n(t_{r(n)} - t_n)|^2 = 0. \tag{3.15}$$

Downloaded 02/23/22 to 222.29.96.30 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

For notational simplicity, let the X_n and Y_j^n be real valued henceforth in the proof. The proof for the general case is the same. We can write

$$E|F^n(t_{r(n)}+\delta-t_n)-F^n(t_{r(n)}-t_n)|^2=\frac{1}{q_n}E\sum_{i,j=r(n)}^{r(n)+q_n\delta}Y_i^nY_j^n. \tag{3.16}$$

By (3.8) and the fact that $\sum_{j=m}^{\infty}e^{-\lambda(t_j-t_m)}a_j=O(1)$ uniformly in m, the above expression equals

$$\frac{1}{q_n}[\delta^2q_n^2O(a_n^{3/2})+\delta q_nO(1)]$$

which goes to zero as needed due to (A3.1).

Part 4. We next show that the limit of any weakly convergent subsequence of $\{F^n(\cdot)\}$ is a martingale. Let $f(\cdot)$ be any bounded and continuous function of its arguments. For any integer p, fix $s\geq 0,\tau\geq 0$, and let $s_i\leq s, i=1,\cdots,p$. Then

$$\begin{aligned}&Ef(F^n(s_i),i\leq p)[F^n(s+\tau)-F^n(s)]\\&\quad=Ef(F^n(s_i),i\leq p)E_{n+q_ns}[F^n(s+\tau)-F^n(s)],\end{aligned}$$

where E_{n+q_ns} is the expectation given all data up to iterate $n+q_ns$ or, equivalently, given all data which is used to calculate $F^n(u),u\leq s$.

We have

$$\begin{aligned}&E|E_{n+q_ns}F^n(s+\tau)-F^n(s)|\\&\quad=E\frac{1}{\sqrt{q_n}}\left|E_{n+q_ns}\sum_{i=n+q_ns}^{n+q_ns+q_n\tau}Y_i^n\right|.\end{aligned}$$

By (3.13) and Lemma 3.1, this expression equals

$$\frac{1}{\sqrt{q_n}}\left[q_nO(a_n)\tau+O(a_n^{1/2})\sum_{i=n+q_ns}^{n+q_ns+q_n\tau}e^{-\lambda(t_i-t_n)}\right] \tag{3.17}$$

$$\leq\sqrt{q_n}O(a_n)\tau+O(1)\sum_{i=n+q_ns}^{\infty}e^{-\lambda(t_i-t_n)}a_i/\sqrt{a_nq_n},$$

which goes to zero as $n\to\infty$ uniformly in any bounded τ-interval. Let $F(\cdot)$ denote the limit of a weakly convergent subsequence of $\{F^n(\cdot)\}$. By the fact that the right side of (3.16) is $O(\delta)$, $\{F^n(t)\}$ is uniformly integrable, for each $t<\infty$. This and the fact that expression (3.17) goes to zero as $n\to\infty$ yields

$$Ef(F(s_i),i\leq p)[F(s+\tau)-F(s)]=0$$

for all s,τ,s_i,p and $f(\cdot)$ in the chosen classes. This implies that $F(\cdot)$ is a martingale. Since the discontinuities in $F^n(\cdot)$ are $O(q_n^{-1/2})$ and tend to zero as $n\to\infty, F(\cdot)$ is continuous. To prove that it is the asserted Wiener process, we need only identify its quadratic variation. This will be done in Part 6. In preparation for that, we need the following uniform integrability result.

Downloaded 02/23/22 to 222.29.96.30 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 02/23/22 to 222.29.96.30 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Part 5. *A uniform integrability result.* For $s > 0, \tau > 0$, and given $T < \infty$ define the sets

$$I_\nu^n = [n + q_n s + \nu T/a_n, n + q_n s + (\nu + 1)T/a_n),$$

$\nu = 0, 1, \cdots, \tau K_n - 1$, (assuming τK_n is an integer without loss of generality) where K_n is defined by

$$K_n(T/a_n) = q_n. \tag{3.18}$$

Then, starting at time s to "cover" $[s, s + \tau)$, we need $K_n\tau$ groups of (T/a_n) iterates each. Define

$$\delta F_\nu^n = \left(\frac{a_n}{T}\right)^{1/2} \sum_{i \in I_\nu^n} Y_i^n.$$

We will show that, for each T,

$$\sup_{\nu \le K_n \tau} E|\delta F_\nu^n|^4 < \infty. \tag{3.19}$$

For simplicity of notation, we work only with the real valued Y_i^n case. For $\nu \le K_n\tau$, there is a $k < \infty$, such that the indices i for which Y_i^n is in δF_ν^n are all in the interval $[n, n + kn]$, so that (2.4) and the ratio $a_n/a_i \approx 1$ can be used. We have

$$E|\delta F_\nu^n|^4 = O(1)\left(\frac{a_n}{T}\right)^2 \sum_{i \le j \le k \le \ell} EY_i^n Y_j^n Y_k^n Y_\ell^n,$$

where the indices vary over the set I_ν^u, subject to the indicated inequalities. By (3.13), this expression can be written as

$$O(1)\left(\frac{a_n}{T}\right)^2 \sum_{i \le j \le k \le \ell} EY_i^n Y_j^n Y_k^n E_k Y_\ell^n$$

$$\le O(1)\left(\frac{a_n}{T}\right)^2 \sum_{i \le j \le k \le \ell} E|Y_i^n Y_j^n Y_k^n|$$

$$\times [e^{-\lambda(t_\ell - t_k)}|Y_k^n| + a_n].$$

Now, using Lemma 3.1 and the fact that $\sum_{j \ge n} e^{-\lambda(t_j - t_n)} a_j = O(1)$, uniformly in n yields that the sum is $O(1)$, uniformly in n and in the range of ν in question, which proves the assertion (3.19).

Part 6. *The quadratic variation of the limit process.* For notational simplicity, we continue to work with the case of real valued X_n, Y_i^n. We first present some consequences of the weak convergence in Theorem 2.1. Recall that

$$\sqrt{\frac{a_n}{T}} \sum_{i \in I_v^n} X_i \to N(0, V_T),$$

Downloaded 02/23/22 to 222.29.96.30 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

where the symbol ($\to N(0, V_T)$) means convergence in distribution to a normally distributed random variable with mean zero and variance V_T. Recall that $V_T = V + O(1/t)$.

Also, by the weak convergence in Theorem 2.1, the set

$$\left\{\left(\frac{a_n}{T}\right)^{1/2} \sum_{i \in I^n_{k_j}} X_i,\ j = 1, \cdots, q\right\}$$

converges in distribution to a set of normal random variables $\{Z_1, \cdots, Z_q\}$, with mean zero and covariance

$$\operatorname{cov}[Z_i, Z_j] = O\left(e^{-\lambda T|k_i - k_j|}\right).$$

Furthermore, there is $\varepsilon(T) \to 0$ as $T \to \infty$ such that the following holds if $N_n \to \infty$ slowly enough:

$$\frac{1}{N_n} \sum_{\nu=1}^{N_n} (\delta F^n_\nu)^2 \xrightarrow{P} V(T), \tag{3.20}$$

where $V(T) \in [V - \varepsilon(T), V + \varepsilon(T)]$ and $\xrightarrow{P}$ denotes convergence in probability. The result is a consequence of the weak convergence in Theorem 2.1, the equivalences and estimates in Part 1 of the proof, the uniform integrability of $\{(\delta F^n_\nu)^2, \nu \le kK_n, n\}$ for any $k < \infty$, and a law of large numbers. By the uniform integrability, (3.20) also holds in the mean.

The bound on ν is chosen to assure that the time indices i of all the iterates Y^n_i fall in the range $[n, n + 2kn]$, so that (2.4) can be used.

Let n index a weakly convergent subsequence of $\{F^n(\cdot)\}$ with limit $F(\cdot)$. Let $f(\cdot)$ be a bounded continuous function, and let $\phi(\cdot)$ be continuous with compact support with the first two derivatives being continuous. For any integer p, let $s_i \le s, i \le p$, and let $\tau > 0$. To get the quadratic variation result, we need to show that for all such $f(\cdot), \phi(\cdot), s_i, \tau, s, p$,

$$\begin{aligned} &Ef(F^n(s_i), i \le p)[\phi(F^n(s+\tau)) - \phi(F^n(s))] \\ &\quad \to Ef(F(s_i), i \le p)[\phi(F(s+\tau)) - \phi(F(s))] \\ &\quad = Ef(F(s_i), i \le p)\left[\frac{V}{2}\int_s^{s+\tau} \phi_{FF}(F(u))du\right]. \end{aligned} \tag{3.21}$$

Note that

$$F^n\left(s + \frac{(\nu+1)T}{a_n q_n}\right) - F^n\left(s + \frac{\nu T}{a_n q_n}\right) = \frac{1}{K_n^{1/2}} \delta F^n_\nu.$$

Now, expanding the left side of (3.21) yields

(3.22)

$$\begin{aligned} &Ef(F^n(s_i), i \le p) \\ &\times\left[\frac{1}{K_n^{1/2}} \sum_{\nu=1}^{K_n\tau} \phi_F\left(F^n\left(s + \frac{vT}{a_n q_n}\right)\right)\delta F^n_v + \frac{1}{2K_n} \sum_{\nu=1}^{K_n\tau} \phi_{FF}\left(F^n\left(s + \frac{\nu T}{a_n q_n}\right)\right)(\delta F^n_\nu)^2 \right. \\ &\qquad \left. + \frac{1}{K_n^{3/2}} \sum_{\nu=1}^{K_n\tau} O(|\delta F^n_\nu|^3)\right]. \end{aligned}$$

Downloaded 02/23/22 to 222.29.96.30 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

By the results in Parts 1, 4, and 5, (3.22) is asymptotically equivalent to

$$Ef(F^n(s_i), i \le p)\left[\frac{1}{2K_n}\sum_{\nu=1}^{K_n\tau}\phi_{FF}\left(F^n\left(s+\frac{\nu T}{a_n q_n}\right)\right)(\delta F_\nu^n)^2\right]. \tag{3.23}$$

By the tightness of $\{F^n(\cdot)\}$ and the uniform integrability of $\{|\delta F_v^n|^2\}$ shown by (3.19), we can "delay" the time argument of the $F^n(\cdot)$ in the ϕ_{FF} in (3.23) by an amount which goes to zero as $n \to \infty$, without changing the asymptotic value. This observation allows us to regroup the summands in (3.23) so that (3.20) can be used. In fact, (3.23) is asymptotically equivalent to (3.24), where we define v_n by $\tau K_n = N_n v_n$, with $N_n \to \infty$ as slowly as we wish but such that $K_n \to \infty$ and $v_n \to \infty$:

$$\begin{aligned} &Ef(F^n(s_i), i \le p) \\ &\quad\times\left[\frac{1}{2v_n}\sum_{\nu=1}^{v_n}\left\{\phi_{FF}\left(F^n\left(s+\frac{\nu N_n T}{a_n q_n}\right)\right)\frac{1}{N_n}\sum_{u=\nu N_n}^{\nu N_n+N_n-1}(\delta F_u^n)^2\right\}\right]. \end{aligned} \tag{3.24}$$

Finally, applying (3.20) to (3.24), taking limits as $n \to \infty$, and using the arbitrariness of T and $\varepsilon(T)$ yields the desired result, namely, the right side of (3.21).

Part 7. Dropping condition (*). When (*) is dropped, we need to regroup certain terms so that the same asymptotic expansions will hold. Define the increasing sequence ρ_v (depending on n) recursively by $\rho_0 = 0$, and for $v \ge 1$

$$\sum_{n+q_n s+\rho_{v-1}}^{n+q_n s+\rho_v} a_j \to T$$

as $n \to \infty$. Redefine the sets of indices I_v^n to be $I_v^n = [n+q_n s+\rho_{v-1}, n+q_n s+\rho_v)$. Thus the sums of the a_i in each set equal T asymptotically. Set $m(n,v) = n + q_n s + \rho_{v-1}$, the first index in the set I_v^n. In sums of the form $(a_n/T)^{1/2}\sum I_v^n$, replace the a_n by $a_{m(n,v)}$. Define $J_n = \min\{\alpha : \rho_\alpha \ge q_n\tau\}$. The J_n replace the K_n in the proof. Finally in the expansions from (3.22) to (3.24), replace the vT/a_n by ρ_v.

With these changes the proof goes through as done above. □

Appendix.

Proof of Lemma 3.1. The proof will be given for the $\{X_n\}$ only. The proof for the $\{Y_i^n\}$ follows from this, and the details are omitted.

Part 1. *Mean square bounds.* A perturbed Liapunov function method will be used. Define the perturbation

$$V_1(x,n) = \sum_{j=n}^{\infty} a_j V_x'(x) E_n\psi_j(x) = O(a_n)|V_x(x)|,$$

where the right-hand inequality is due to (A3.2). We can write

$$\begin{aligned} E_n V(X_{n+1}) - V(X_n) &= a_n V_x'(X_n)\bar{g}(X_n) \\ &\quad + a_n V_x'(X_n)\psi_n(X_n) + a_n^2 O(1)|g(X_n,\xi_n)|^2 \end{aligned} \tag{1}$$

$$\begin{aligned} &E_n V_1(X_{n+1}, n+1) - V_1(X_n, n) \\ &= -a_n V_x'(X_n)\psi_n(X_n) + \sum_{j=n+1}^{\infty} E_n a_j [V_x'(X_{n+1})\psi_j(X_{n+1}) - V_x'(X_n)\psi_j(X_n)]. \end{aligned} \tag{2}$$

Downloaded 02/23/22 to 222.29.96.30 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

The fact that the second term on the right side of (1) is the negative of the first term on the right side of (2) is the essential motivation behind the construction of $V_1(\cdot)$. Rewriting the last term on the right side of (2) as the sum of the left-hand sides in (3a), (3b) following and bounding them by use of (A3.2)–(A3.5) yields:

$$\text{(3a)} \qquad \left|\sum_{n+1}^{\infty} E_n a_j V_x'(X_n)(\psi_j(X_{n+1}) - \psi_j(X_n))\right|$$

$$\leq |V_x(X_n)|O(a_n^2)[V^{1/2}(X_n) + 1] \leq O(a_n^2)[V(X_n) + 1],$$

$$\text{(3b)} \qquad \left|\sum_{n+1}^{\infty} a_j E_n[V_x'(X_{n+1}) - V_x'(X_n)]\psi_j(X_{n+1})\right|$$

$$\leq O(a_n^2)|g(X_n, \xi_n)| \leq O(a_n^2)[V(X_n) + 1].$$

Define the perturbed Liapunov function $\tilde{V}_n = V(X_n) + V_1(X_n, n)$. Putting the estimates (2) and (3) together yields

$$E_n\tilde{V}_{n+1} - \tilde{V}_n \leq -a_n\gamma V(X_n) + O(a_n^2)$$

$$+O(a_n^2)[1 + V(X_n)]$$

and

$$\text{(4)} \qquad E_n\tilde{V}_{n+1} - \tilde{V}_n \leq -\frac{a_n}{2}\gamma\tilde{V}_n + O(a_n^2).$$

Equation (4) implies that $\{E\tilde{V}_n/a_n, n < \infty\}$ is bounded from above. Then, using this and the estimate $V_1(x, n) = O(a_n)[V(x) + 1]$ yields the boundedness of $\{EV(X_n)/a_n, n < \infty\}$. This latter bound and the first equation of (A3.5) yield the boundedness of $\{E|X_n|^2/a_n, n < \infty\}$.

Equation (4) implies that there are β_n such that $E|\beta_n| < \infty, E_n\beta_n = 0$, and

$$\text{(5)} \qquad \tilde{V}_{n+1} - \tilde{V}_n \leq \frac{-a_n}{2}\gamma\tilde{V}_n + O(a_n^2) + \beta_n,$$

from which we get

$$E_n\tilde{V}_q = O(a_n)$$

for any $\mathcal{F}_m$-stopping time q with values in $[n, n + kn]$. This, together with the above given bound on $V_1(x, n)$, yields the second assertion of the lemma.

Part 2. Fourth moments. We now prove $E|X_n|^4 = O(a_n^2)$. The procedure will be similar to that used in Part 1. Define the perturbation

$$\text{(6)} \qquad V_2(x, n) = 2\sum_{j=n}^{\infty} a_j E_n V(x)V_x'(x)\psi_j(x) = O(a_n)|V(x)V_x(x)|,$$

and the perturbed Liapunov function

$$\text{(7)} \qquad \hat{V}(x, n) = V^2(x) + V_2(x, n).$$

We use $\tilde{X}_n$ and $\hat{X}_n$ to denote vectors in the interval $[X_n, X_{n+1}]$, and their values might change from case to case. Proceeding as for the second-order case, by a truncated Taylor series expansion we can write

$$E_n V^2(X_{n+1}) - V^2(X_n) = B_1 + B_2 + B_3 + B_4, \tag{8}$$

where

$$B_1 = 2a_n V(X_n) V_x'(X_n) \bar{g}(X_n),$$

$$B_2 = 2a_n E_n V(X_n) V_x'(X_n) \psi_n(X_n),$$

$$B_3 = a_n^2 E_n g'(X_n, \xi_n) V_x(\tilde{X}_n) V_x'(\tilde{X}_n) g(X_n, \xi_n),$$

$$B_4 = a_n^2 E_n V(\tilde{X}_n) g'(X_n, \xi_n) V_{xx}(\tilde{X}_n) g(X_n, \xi_n).$$

Furthermore,

$$E_n V_2(X_{n+1}, n+1) - V_2(X_n, n) = B_5 + B_6, \tag{9}$$

where

$$B_5 = -2a_n E_n V(X_n) V_x'(X_n) \psi_n(X_n),$$

$$B_6 = 2 \sum_{j=n+1}^{\infty} a_j E_n \Big[V(X_{n+1}) V_x'(X_{n+1}) \psi_j(X_{n+1}) - V(X_n) V_x'(X_n) \psi_j(X_n) \Big].$$

The terms $B_i, i = 1, \cdots, 6$, will now be bounded. Heavy use will be made of the inequalities in (A3.5) and the fact that $EV(X_n) = O(a_n)$ by Part 1 of the proof. We have

$$B_1 \leq -2a_n \gamma V^2(X_n).$$

B_2 is cancelled by B_5. For appropriate $\tilde{X}_n$ and $\hat{X}_n$,

$$\begin{aligned} B_3 &= a_n^2 E_n g'(X_n, \xi_n) [V_x(X_n) + V_{xx}(\hat{X}_n)(\tilde{X}_n - X_n)] \\ &\quad \times [V_x(X_n) + V_{xx}(\hat{X}_n)(\tilde{X}_n - X_n)]' g(X_n, \xi_n) \\ &= C_1 + C_2 + C_3, \end{aligned}$$

where

$$EC_1 = 2a_n^2 E g'(X_n, \xi_n) V_x(X_n) V_x'(X_n) g(X_n, \xi_n)$$

$$EC_2 = O(a_n^2) E |g(X_n, \xi_n)|^2 |V_x(X_n)| |\tilde{X}_n - X_n|$$

$$EC_3 = O(a_n^2) E |g(X_n, \xi_n)|^2 |\tilde{X}_n - X_n|^2.$$

Downloaded 02/23/22 to 222.29.96.30 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 02/23/22 to 222.29.96.30 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

We have the following bounds:

$$\begin{aligned} EC_1 &\le O(a_n^2)EV(X_n)(V(X_n)+1) \\ &\le O(a_n^2)EV^2(X_n) + O(a_n^3) \\ EC_2 &\le O(a_n^2)E(V(X_n)+1)|V_x(X_n)||\tilde{X}_n - X_n| \\ &\le O(a_n^2)E(V(X_n)+1)|V_x(X_n)||X_{n+1} - X_n| \\ &\le O(a_n^3)E(V(X_n)+1)^2 \le O(a_n^3)EV^2(X_n) + O(a_n^3), \\ EC_3 &\le O(a_n^2)E(V(X_n)+1)a_n^2(V(X_n)+1) \\ &\le O(a_n^4)EV^2(X_n) + O(a_n^4). \end{aligned}$$

By a similar method, the other terms can be shown to satisfy

$$EB_i = O(a_n^2)EV^2(X_n) + O(a_n^3), \qquad i = 4,5,6.$$

Finally, putting these estimates together yields

$$\begin{aligned} E\hat{V}(X_{n+1}, n+1) - E\hat{V}(X_n) \le &-\gamma a_n EV^2(X_n) \\ &+ O(a_n^2)EV^2(X_n) + O(a_n^3). \end{aligned} \tag{10}$$

Note that

$$\begin{aligned} |V_2(X_n, n)| &= O(a_n)V(X_n)(V(X_n)+1) \\ &\le O(a_n)EV^2(X_n) + O(a_n^2). \end{aligned} \tag{11}$$

Using (10) and (11) yields, for large n,

$$E\hat{V}(X_{n+1}, n+1) - E\hat{V}(X_n, n) \le -\frac{\gamma}{2}a_n E\hat{V}(X_n, n) + O(a_n^3). \tag{12}$$

Then, following the procedure of Part 1, we get

$$\sup_n EV^2(X_n)/a_n^2 < \infty$$

and

$$E|X_n|^4/a_n^2 < \infty. \qquad \square$$

REFERENCES

[1] B. T. POLYAK, *New stochastic approximation type procedures*, Automat. i Telemekh., 7 (1990), pp. 98–107.

[2] B. T. POLYAK AND A. B. JUDITSKY, *Acceleration of stochastic approximation by averaging*, SIAM J. Control Optim., 30 (1992), pp. 838–855.

[3] D. RUPPERT, *Efficient estimators from a slowly convergent Robbins–Monro process*, Tech. Report, No. 781, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 1988.

[4] G. YIN, *Stochastic approximation via averaging; Polyak's approach revisited*, Lecture Notes in Economics and Mathematical Systems 374, G. Pflug and U. Dieter, eds. Springer-Verlag, Berlin, 1992, pp. 119–134.

[5] G. YIN, *On extensions of Polyak's averaging approach to stochastic approximation*, Stochastics, 36 (1992), pp. 245–264.
[6] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.
[7] S. N. ETHIER AND T. G. KURTZ, *Markov Processes: Characterization and Convergence*, John Wiley, New York, 1986.
[8] H. J. KUSHNER AND D. S. CLARK, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Applied Math. Sciences Series, Springer-Verlag, Berlin, New York, 1978.
[9] A. BENVENISTE, M. METIVIER, AND P. PRIORET, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, Berlin, New York, 1990. (Translated from the French.)
[10] H. J. KUSHNER, *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, Birkhauser, Boston, 1990.
[11] H. J. KUSHNER AND HAI HUANG, *Averaging methods for the asymptotic analysis of learning and adaptive systems with small adjustment rate*, SIAM J. Control Optim., 19 (1981), pp. 635–650.
[12] ———, *Approximation and Weak Convergence Methods for Stochastic Processes with Applications to Stochastic Systems Theory*, M.I.T. Press, Cambridge, 1984.
[13] H. J. KUSHNER AND G. YIN, *Asymptotic properties of distributed and communicating stochastic approximation algorithms*, SIAM J. Control Optim., 25 (1987), pp. 1266–1290.
[14] H. J. KUSHNER AND A. SHWARTZ, *An invariant measure approach to the convergence of stochastic approximations with state dependent noise*, SIAM J. Control Optim., 22 (1984), pp. 13–27.
[15] H. J. KUSHNER, *Stochastic approximation with discontinuous dynamics and state dependent noise*, J. Math. Anal. Appl., 82 (1981), pp. 527–542.

Downloaded 02/23/22 to 222.29.96.30 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy