Journal of Computational Mathematics, Vol.8, No.1, 1990, 45–54.

CONVERGENCE SPEED AND ASYMPTOTIC DISTRIBUTION OF A PARALLEL ROBBINS–MONRO METHOD*

Zhu Yun-min[1)]
(*Institute of Mathematical Sciences, Chengdu Branch, Academia Sinica, Chengdu, China*)
Yin Gang[2)]
(*Department of Mathematics, Wayne State University, Detroit, USA*)

Abstract

Very recently, there is a growing interest in studying parallel and distributed stochastic approximation algorithms. Previously, we suggested such an algorithm to find zeros or locate maximum values of a regression function with large state space dimension in [1], and derived the strong consistency property for that algorithm. In the present work, we concern ourselves with the problem of asymptotic properties of such an algorithm. We will study the limit behavior of the algorithm and obtain the rate of convergence and asymptotic normality results.

§1. Introduction

Very recently, there is a growing interest in studying parallel and distributed stochastic approximation algorithms. [1]–[5] proposed several such schemes. The purposes of the studies are to exploit the opportunities provided by parallel processing methods and take advantage of the asynchronous communication.

Motivated by [5], we suggested a parallel stochastic approximation algorithm in [1] to locate seros or maximum values of a regression function with large state space dimension. The methods of random truncations were employed in order to obtain the boundedness of the algorithm and to enssure the convergence. By the truncation techniques, we were able to treat a rather broad class of regression functions. The strong consistency property for the aforementioned algorithm was proved under rather weak conditions.

To study any kind of recursive algorithm, there are basically two questions that one wants to answer. First, one wants to see if the algorithm works (convergence); next, if the algorithm converges, one would like to find the convergence speed. In this paper, we will concern ourselves with the second problem, namely rate of convergence.

The algorithm suggested in [1] is a generalization of the relaxation method. Such an idea was originally given in [5]. There are two distinct features for the parallel RM algorithm proposed in [1]. First, there is no iteration number which is a common index to all the processors. Second, the computation intervals are random. Intuitively, we would expect that similar 'rate' of convergence result for the classical algorithm still holds for the parallel algorithms. However, because of the asynchronization and additional randomness (random computation times) coming in, the analysis is not straightforward. Since we must take account of available information for all the processors, the notations as well as the analysis are pretty complicated. One of the key points here is to overcome the difficulties of different

* Received May 25, 1987.
1) Research of this author was supported in part by Science Foundation of Academia Sinica.
2) Research of this author was supported in part by Wayne State University under the Wayne State University Research Award.

computation times for different processors. When evaluating the limit, we thus need to work on each processor separately.

The paper is organized as follows. In Section 2, the basic formulation and some conditions are stated. In Section 3, the result of $n^\delta x_n \to 0$ for some δ, with $0 < \delta < \frac{1}{2}$, is obtained. Finally, in Section 4, some discussion on the asymptotic normality is derived.

§2. The Algorithm

Let x be an r-dimensional vector, $x = (x^1, \cdots, x^r)'$. Let there be r processors, each controlling one component of the state vector. For each $i \le r$, let processor i takes y_j^i units of time to complete the jth iteration, where $\{y_j^i\}$ is a sequence of positive integer valued random variables which may depend on state and noise. Define τ_n^i by

$$\tau_0^i = 0, \quad \tau_n^i = \sum_{j=1}^{n} y_j^i.$$

$\{\tau_n^i\}$ is the random computation time. It is quite similar to the conventional renewal process. Let $\xi_{\tau_n^i}^i$ be the noise incurred in the nth iteration for x^i. Let $x_1 = (x_1^1, \cdots, x_1^r)'$ be the initial value and $x_{\tau_n^i}^i$ be the value of the ith component of the state at the end of the nth iteration (or nth processing.time). For $n \in [\tau_j^i, \tau_{j+1}^i)$, put

$$x_n^i = x_{\tau_j^i}^i,$$
$$\xi_n^i = \xi_{\tau_j^i}^i,$$
$$x_{\tau_n^i} = (x_{\tau_n^i}^1, \cdots, x_{\tau_n^i}^r)'.$$

Let $b(\cdot) : R^r \longmapsto R^r$ be a continuous function, $b(\cdot) = (b^1(\cdot), \cdots, b^r(\cdot))'$ and $\varepsilon_n = \frac{1}{n}$; the basic algorithm to be considered is

$$x_{\tau_{n+1}^i}^i = x_{\tau_n^i}^i + \varepsilon_{\tau_n^i}^i (b^i(x_{\tau_n^i}) + \xi_{\tau_n^i}^i), \quad i \le r. \tag{2.1}$$

Remark. Comparing with the classical RM algorithm here we emphasize parallel aspects of our algorithm. Starting with initial value, new values of x^i are computed based on the most recently determined values of x^j, for $j \le r$. The newly computed values are passed to all other components of x. This is a generalization of the relaxation methods.

Since each processor takes a random time to complete each iteration, and this random time in general is different from processor to processor, there is no "iteration number" which is a common index for all the processors. This causes difficulties in both notation and analysis, and we have to use elapsed processing time (real time) as the time indicator.

To assume each processor to control only one component of the state vector is really no loss in generality. In fact, we could consider the case that each processor controls several components of the state vector. The notation would be more complicated, but the analysis essentially remains the same.

To proceed, we also need the following definitions.

$$\begin{aligned} N_i(n) &= \sup\{k; \tau_k^i \le n\}, \\ \Delta_n^i &= n - \tau_{N_i(n)}^i, \\ I_n^i &= I_{\{\Delta_n^i = 0\}}. \end{aligned} \tag{2.2}$$

In the case of usual renewal process, $N_i(n)$ is the counting process. It counts the number of events or renewals up to time n. Here it serves the same purpose of counting the number of iterations for component i up to time n. Δ_n^i is the so–called "age" or "current life" process which represents the time that has elapsed since last iteration. If $\Delta_n^i = 0$, then n is a random computation time.

To keep track of all the state values used for the r processors at each time n, we define an augmented vector, $\tilde{x}$, such that $\tilde{x}_n = (\tilde{x}_n^1, \cdots, \tilde{x}_n^r)'$, where for each $i \leq r$,

$$\tilde{x}_n^i = x_{r^i_{N_i(n)}} = \left(x^1_{r^i_{N_i(n)}}, \cdots, x^r_{r^i_{N_i(n)}}\right).$$

With these notations, we can rewrite (2.1) as

$$x_{n+1}^i = x_n^i + \varepsilon_n(b^i(\tilde{x}_n^i) + \xi_n^i)I_{n+1}^i, \quad i \leq r \tag{2.3}$$

or in a vector form

$$x_{n+1} = x_n + \varepsilon_n I_{n+1}(b(\tilde{x}_n) + \xi_n) \tag{2.4}$$

where

$$x_n = (x_n^1, \cdots x_n^r)',$$

$$\xi_n = (\xi_n^1, \cdots, \xi_n^r)',$$

$$b(\tilde{x}_n) = (b^1(\tilde{x}_n^1), \cdots, b^r(\tilde{x}_n^r))',$$

$$I_n = \text{ diag } (I_n^1, \cdots, I_n^r).$$

Remark. For each n, if $I_{n+1}^i = 0$, no update action is taken, $x_{n+1}^i = x_n^i$, and the state keeps its old value without any change. If $I_{n+1}^i = 1$, then $x_{n+1}^i = x_n^i + \varepsilon_n(b^i(\tilde{x}_n^i) + \xi_n^i)$, and one step of update is performed.

The state values are communicated to all other components as soon as they are available.

Let $b^i(x) = 0$ have a unique root, say x_0, for each $i \leq r$. Suppose we have a twice continnously differentiable function $v(\cdot)$, such that

(i) $$v(x) \neq v(x_0) \quad \forall\, x \neq x_0,$$

(ii) $$v_x(x)\mu^{-1}b(x) < 0 \quad \forall\, x \neq x_0$$

where μ^{-1} is given by

$$\mu^{-1} = \text{ diag}\left(\frac{1}{\mu^1}, \cdots, \frac{1}{\mu^r}\right),$$

(iii) $$|v(x)| \to \infty, \quad \text{as} \quad |x| \to \infty.$$

Let $\{M_n\}$ be a sequence of monotone increasing positive real numbers, such that $M_n \to \infty$ as $n \to \infty$. Define σ_n as

$$\sigma_0 = 0,$$

$$\sigma_{n+1} = \sigma_n + I_{\{|x_n + \frac{1}{n}(b(\tilde{x}_n)+\xi_n)| > M_{\sigma_n}\}},$$

$$J_n = I_{\{|x_n + \frac{1}{n}(b(\tilde{x}_n)+\xi_n)| \leq M_{\sigma_n}\}},$$

$$J_n^c = I_{\{|x_n + \frac{1}{n}(b(\tilde{x}_n)+\xi_n)| > M_{\sigma_n}\}}.$$

Define

$$x_{n+1} = [x_n + \frac{1}{n} I_{n+1}(b(\tilde{x}_n) + \xi_n)]J_n + \hat{x}J_n^c. \tag{2.5}$$

Now the paragraph after (2.4) and the following conditions (A1), (A2) imply that the sequence $\{x_n\}$ given by equation (2.5) is strongly consistent, i.e. $x_n \to x_0$ w.p.l. For a detailed proof, the readers are referred to [1]. Without loss of generality, in the sequel, we will assume that $x_0 = 0$.

(A1) For each $i \le r$, $\{y_n^i\}$ is bounded and

$$y_n^i = y_n^i(x^i_{r^i_{n-1}}, \xi^i_{r^i_{n-1}}),$$

$$E(y_n^i - \mu_n^i | y_l^i - \mu_l^i) = 0, \quad l < n$$

where μ_n^i is a sequence of random variables satisfying

$$\mu_n^i \xrightarrow{n} \mu^i, \quad \text{a.s.}$$

with μ^i a constant.

$$\sum_{n=1}^{\infty} n^{\frac{1}{2}(1-\alpha)}(\mu_n^i - \mu_{n-1}^i)$$

converges a.s. and

$$E(y_n^i - \mu_n^i)^2 \le M$$

for some α, with $0 < \alpha < 1$.

(A2) For each $i < r$, the noise $\xi^i_{r^i_n}$ satisfies

$$\xi^i_{r^i_n} = \phi^i_{r^i_n} + \psi^i_{r^i_n},$$

$$(\tau_n^i)^\delta \psi^i_{r^i_n} \xrightarrow{n} 0, \quad \text{a.s.}$$

$$\sum \varepsilon_{r^i_j} \phi^i_{r^i_j} \text{ converges a.s.}$$

(A3) $b(x) = Dx + o(|x|)$ where D is an $r \times r$ real symmetric negative definite matrix. Let $-\lambda_i$ denote the eigenvalues of D, with $\lambda_i > 0$. Let $\rho = \min_{1 \le i \le r} \lambda_i$ and let $\tilde{M}$ be the uniform bound for $\{y_j^i\}$, i.e.

$$\tilde{M} = \max_i \sup_j |y_j^i| \quad \text{such that} \quad \rho > \frac{1}{2}\tilde{M}.$$

Remark. Assumption (A3) says that $b(\cdot)$ consists of a linear part and a the high order nonlinear part (high order w.r.t. x). This is a standard assumption in proving the rate of convergence result. Although x_n and $\tilde{x}_n$ are not the same vectors, as n is getting larger and larger, the difference between the two is getting smaller and smaller because of the fact that $x_n \to 0$ as $n \to \infty$ (see the paragraph after (2.4)). Since in the computation we have to use the vector $\tilde{x}_n$, we define a block diagonal matrix $\tilde{D}$ by

$$\tilde{D} = \text{ diag } (D_1, \cdots, D_r)$$

with D_i equal to the ith row of the matrix D. It follows from (A3) that $b(\tilde{x}_n) = \tilde{D}\tilde{x}_n + o(|\tilde{x}_n|)$.

The assumption on D is a stability condition. The requirement for D to be negative definite is not essential. As can be seen in the sequel, all subsequent development can be

extended to the case that D is a stable matrix, i.e., all eigenvalues of D have negative real parts. In order to make the presentation clear, we choose the stronger condition (A3) as in the present form.

Since y_n^i is a sequence of bounded random variables, the assumption $\rho > \frac{1}{2}\tilde{M}$ implies that $\rho > \delta y_j^i$ for all $0 < \delta < \frac{1}{2}$, all j and $i \leq r$.

As we remarked in [1], we choose the gain sequence to be of the particular form for purely notational simplicity. Other forms of the gain sequence can be considered. However, the idea here is to investigate the convergence and convergence rate of the parallel RM like algorithm with random truncations, not to find the most general gain sequence.

§3. $x_n = o(n^{-\delta})$

It is fairly easy to see that the convergence speed for the parallel RM like algorithm with randomly varying truncations is the same as that of the algorithm without truncations. So we need to consider only the latter one.

To start with, using (A3), write equation (2.3) as

$$x_{n+1} = x_n + \frac{1}{n} I_{n+1}(\tilde{D}\tilde{x}_n + o(|\tilde{x}_n|) + \xi_n) \tag{3.1}$$

and denote $\hat{x}_n = (x_n, \cdots, x_n)'$, i.e. $\hat{x}_n$ is an $r \times r$ vector with each subdivision equal to x_n. We have

$$\begin{aligned} x_{n+1} &= x_n + \frac{1}{n} I_{n+1}(\tilde{D}\tilde{x}_n + \xi_n) + \frac{1}{n} I_{n+1}(o|\tilde{x}_n|) + \tilde{D}(\tilde{x}_n - \hat{x}_n) \\ &= x_n + \frac{1}{n} I_{n+1}(Dx_n + o(|x_n|) + \xi_n) + \frac{1}{n} I_{n+1}(o|\tilde{x}_n - \hat{x}_n|) + \tilde{D}(\tilde{x}_n - \hat{x}_n)). \end{aligned} \tag{3.2}$$

Lemma 1. *The last term on the right hand side of* (3.2) *can be written as* $\frac{1}{n} I_{n+1}(B_n + C_n)$, *such that* $n^{\delta} B_n \to 0$ *and* $\sum_{n=1}^{\infty} n^{\delta-1} C_n$ *converges* a.s., *for* $\delta \in (0, \frac{1}{2})$.

Proof. To verify this assertion, we look at the difference

$$\hat{x}_n - \tilde{x}_n = \begin{pmatrix} x_n - \tilde{x}_n^1 \\ \vdots \\ x_n - \tilde{x}_n^r \end{pmatrix}.$$

Recalling the definition of $\tilde{x}_n^i$, we have for each $i \leq r$

$$x_n - \tilde{x}_n^i = \sum_{k=\tau_{N_i(n)}^i}^{n-1} \frac{1}{k} I_{k+1}(b(\tilde{x}_k) + \xi_k). \tag{3.3}$$

Note the summation is over finitely many terms due to the boundedness of the random computation intervals. $x_n \to 0$ a.s. implies $\tilde{x}_n^i \to 0$ a.s. for each $i \leq r$. Also from condition (A2), we know that $\xi_n = \phi_n + \psi_n$, such that $\sum_{n=1}^{\infty} n^{\delta-1}\phi_n$ converges a.s., and $n^{\delta}\psi_n \to 0$ a.s., where ϕ_n, ψ_n are defined in an obvious manner. The lemma thus follows. As a consequence,

$$x_{n+1} = x_n + \frac{1}{n} I_{n+1}(Dx_n + o(|x_n|) + \xi_n) + \frac{1}{n} I_{n+1}(B_n + C_n). \tag{3.4}$$

There is no loss in generality to assume that D is a diagonal matrix, $D = \text{diag}\ (-\lambda_1, \cdots, -\lambda_r)$ with each $\lambda_i > 0$. If D is not a diagonal matrix, then we can choose a nonsingular matrix P (in fact an orthogonal matrix), such that

$$P^{-1}DP = \Lambda = \text{diag}(-\lambda_1, \cdots, -\lambda_r),$$

$$z_n = P^{-1}x_n, \quad \hat{I}_n = P^{-1}I_nP, \quad \xi_n = P^{-1}\xi_n, \quad \hat{B}_n = P^{-1}B_n, \quad \hat{C}_n = P^{-1}C_n.$$

Then (3.2) becomes

$$z_{n+1} = z_n + \frac{1}{n}\hat{I}_{n+1}(\Lambda z_n + o(|z_n|) + \hat{\xi}_n) + \frac{1}{n}I_{n+1}(\hat{B}_n + \hat{C}_n). \tag{3.5}$$

Because the matrix P is invertible, the convergence properties of (3.5) will be exactly the same as that of (3.4). Henceforth, we will let D be a diagonal matrix.

As we commented before, due to the different random computation times for different components, we would be better off to work on each component. This leads to

$$x^i_{\tau^i_{n+1}} = x^i_{\tau^i_n} + \frac{1}{\tau^i_n}(-\lambda_i x^i_{\tau^i_n} + o(|x_{\tau^i_n}|) + \xi^i_{\tau^i_n}) + \frac{1}{\tau^i_n}(B^i_{\tau^i_n} + C^i_{\tau^i_n}). \tag{3.6}$$

Our ultimate goal is to show that $n^\delta x_n \to 0$, as $n \to \infty$. In order to obtain this, we need to show only $n^\delta x^i_n \to 0$ for each $i \le r$. By the following lemma, this reduces to showing that $(\tau^i_n)^\delta \hat{x}^i_{\tau^i_n} \to 0$ (recall x^i_n are the piecewise constant interpolation of $x^i_{\tau^i_n}$).

Lemma 2. *Under* (A1), *if* $(\tau^i_n)^\delta x^i_{\tau^i_n} \to 0$ *as* $n \to \infty$, *then* $n^\delta x^i_n \to 0$ *as* $n \to \infty$, *for each* $i \le r$.

Proof. To prove this lemma, we recall Lemma 3.1 and its corollary in [1], which states that

$$\frac{1}{n}\tau^i_n \to \mu_i,$$

$$\frac{N_i(n)}{n} \to \frac{1}{\mu_i}.$$

With the help of this lemma, we have

$$\lim_n n^\delta x^i_n = \lim_n (\frac{n}{N_i(n)})^\delta (\frac{N_i(n)}{\tau^i_{N_i(n)}})^\delta ((\tau^i_{N_i(n)})^\delta x^i_{\tau^i_{N_i(n)}}) = \lim_n (\tau^i_{N_i(n)})^\delta x^i_{\tau^i_{N_i(n)}} = 0.$$

This proves the lemma.

With Lemma 1 in mind, we now establish the following lemma.

Lemma 3. *Under assumptions* (A1)– (A3), $(\tau^i_n)^\delta x^i_{\tau^i_n} \to 0$ *as* $n \to \infty$.

Proof. Set $v_k = k^\delta x^i_k$. From (3.6) we get

$$v^i_{\tau^i_{n+1}} = (1 + \frac{y^i_n}{\tau^i_n})^\delta [v^i_{\tau^i_n} - \frac{1}{\tau^i_n}\lambda_i v^i_{\tau^i_n} + \frac{1}{(\tau^i_n)^{1-\delta}}(\xi^i_{\tau^i_n} + B^i_{\tau^i_n} + C^i_{\tau^i_n})] + \frac{1}{\tau^i_n}o(|v_{\tau^i_n}|). \tag{3.7}$$

Now (A3) implies that we can write (3.7) as

$$\begin{aligned} v^i_{\tau^i_{n+1}} &= v^i_{\tau^i_n} + \frac{1}{\tau^i_n}[(d_i + o(\frac{1}{\tau^i_n}))v^i_{\tau^i_n} + (\tau^i_n)^\delta \phi^i_{\tau^i_n} + (\tau^i_n)^\delta \psi^i_{\tau^i_n} \\ &\quad + O(\frac{1}{(\tau^i_n)^{1-\delta}})\xi^i_{\tau^i_n} + o(|v_{\tau^i_n}|)] + (1 + \frac{y^i_n}{\tau^i_n})^\delta \frac{1}{(\tau^i_n)^{1-\delta}}(B^i_{\tau^i_n} + C^i_{\tau^i_n}) \end{aligned} \tag{3.8}$$

where $d_i \leq 0$. Note the convergence of $\frac{1}{n}\tau_n^i$ implies $\sum_n \frac{1}{\tau_n^i} = \infty$ or each $i \leq r$, and by assumption $(\tau_n^i)^\delta \hat{\xi}_{\tau_n^i}^i \to 0$ and

$$O\Big(\frac{1}{(\tau_n^i)^{1-\delta}}\xi_{\tau_n^i}^i\Big) \to 0.$$

The result in [6] together with (A1)–(A3) now yield that $v_{\tau_n^i}^i \xrightarrow{n} 0$. Thus we obtain the following theorem.

Theorem 1. *If* (A1)–(A3) *hold, then* $x_n = o(n^{-\delta})$.

Actually, the theorem says more than one can see at first glance; since if $x_n = o(n^{-\delta})$, for some $0 < \delta < \frac{1}{2}$, then $n^{\delta'} x_n \to 0$ for all $0 < \delta' < \delta < \frac{1}{2}$.

§4. Asymptotic Normality

In [1], we proved that the parallel RM algorithm with randomly varying truncations would become a parallel RM algorithm with bounded iterates after finitely many iterations, i.e. the algorithm is uniformly bounded for $n > M$, for some M. So for any $\eta > 0$, there is an $M_\sigma > 0$, such that

$$P\{\sup_{n \geq M} |x_n| \leq M_\sigma\} > 1 - \eta, \quad \text{where} \quad \sigma = \lim_n \sigma_n.$$

Let $x_n, \tilde{x}_n$ denote the sequence given by the parallel RM algorithm and the parallel RM algorithm with randomly varying truncations respectively. Then

$$P\{\sup_{n \geq M} |x_n - \tilde{x}_n| > 0\} < \eta$$

and for any Borel set B, we have

$$\begin{aligned} &P\{\sqrt{n}x_n \in B\} \\ &\leq P\{\sqrt{n}x_n \in B, \sup_{n \geq M} |x_n - \tilde{x}_n| = 0\} + P\{\sqrt{n}x_n \in B, \sup_{n \geq M} |x_n - \tilde{x}_n| > 0\}. \end{aligned}$$

It follows that

$$\liminf_n P\{\sqrt{n}x_n \in B\} \leq \liminf_n P\{\sqrt{n}\tilde{x}_n \in B\} + \eta.$$

Similarly,

$$\limsup_n P\{\sqrt{n}x_n \in B\} + \eta \geq \limsup_n P\{\sqrt{n}\tilde{x}_n \in B\}.$$

The arbitrariness of η then implies that

$$\lim_n P\{\sqrt{n}x_n \in B\} = \lim_n P\{\sqrt{n}\tilde{x}_n \in B\}.$$

Therefore, $\sqrt{n}x_n$ and $\sqrt{n}\tilde{x}_n$ have the same limit distribution. In order to prove the result for the convergence rate and asymptotic normality, we need only to consider the limit distribution for the parallel RM algorithm. We thus devote our attention to

$$x_{n+1} = x_n + \frac{1}{n} I_{n+1}(b(\tilde{x}_n) + \xi_n). \tag{4.1}$$

We have proved that $n^{\delta}x_n \to 0$. The question now is what happens to the sequence $\sqrt{n}x_n$. In fact, under some suitable conditions this sequence is asymptotically normally distributed. The condition essentially involves assuming ξ^i, ξ^j are independent for $i \neq j$, and the correlations for $\{\xi_n^i\}$ are weak. For notational simplicity, we shall derive the result for D being a diagonal matrix first. For this reason, we require an extra condition (A6). Then in Theorem 3, we relax this condition, and give the result for the case that D is a symmetric negative definite matrix.

(A4) ξ^i and ξ^j are independent for $i \neq j$.

(A5) For each $i \leq r$, $\{y_n^i\}$ do not depend on the state or noise

$$\psi^i_{\tau^i_n} = o\Big(\frac{1}{\sqrt{\tau^i_n}}\Big),$$

$$\phi^i_{\tau^i_n} = \sum_{j=0}^{L} c_j \gamma^i_{\tau^i_{n-j}}, \quad \text{for some } L$$

where $\gamma^i_{\tau^i_k}$ are independent random variables, such that $E\gamma^i_{\tau^i_k} = 0, E(\gamma^i_{\tau^i_k})^2 = r^i, \forall k$, and for any constant K,

$$E|\gamma^i_{\tau^i_k}|^2 I_{(|\gamma^i_{\tau^i_k}| > K\sqrt{\tau^i_k})} \xrightarrow{k} 0.$$

(A6) $D = \text{diag}\,(-\lambda, \cdots, -\lambda_r)$.

We will work on each component again. Since the asymptotic normality was well studied [5], [7], [8], here we make no attempt to spell out the details of the proof; rather, we will make use of the previous results as much as possible.

Define $u_n = \sqrt{n}x_n$; similarly for their components. From the early discussion, we know that

$$u^i_{\tau^i_{n+1}} = (1 + \frac{y^i_n}{\tau^i_n})^{\frac{1}{2}}\{[(1 - \frac{\lambda_i}{\tau^i_n})u^i_n + \frac{1}{\sqrt{\tau^i_n}}\xi^i_{\tau^i_n}] + \frac{1}{\tau^i_n}o(|u_{\tau^i_n}|) + \frac{1}{\sqrt{\tau^i_n}}(B^i_{\tau^i_n} + C^i_{\tau^i_n})\}. \tag{4.2}$$

As a matter of fact, the last two terms inside the curly brackets do not contribute anything to the limit. We need only consider the term

$$(1 + \frac{y^i_n}{\tau^i_n})^{\frac{1}{2}}\{(1 - \frac{\lambda^i}{\tau^i_n})u^i_{\tau^i_n} + \frac{1}{\sqrt{\tau^i_n}}\xi^i_{\tau^i_n}\}. \tag{4.3}$$

Define

$$\alpha^i_{\tau^i_n} = (1 + \frac{y^i_n}{\tau^i_n})^{\frac{1}{2}}, \quad \Phi^i_{\tau^i_n} = \alpha^i_{\tau^i_n}(1 - \frac{\lambda^i}{\tau^i_n}),$$

$$\Phi^i_{n|k} = \prod_{j=k+1}^{n} \Phi^i_{\tau^i_n}, \quad \Phi^i_{n|n} = 1.$$

Then (4.3) can be written as

$$\Phi^i_{\tau^i_n} u^i_{\tau^i_n} + \frac{1}{\sqrt{\tau^i_n}}\alpha^i_{\tau^i_n}\xi^i_{\tau^i_n}, \quad \text{for} \quad i \leq r. \tag{4.4}$$

It follows that

$$u_{n+1}^i = \Phi_{N_i(n)|1}^i u_1^i + \sum_{j=1}^{N_i(n)} \frac{\alpha_{\tau_j^i}^i}{\sqrt{\tau_j^i}} \Phi_{N_i(n)|j}^i \xi_{\tau_j^i}^i + o(1) \tag{4.5}$$

where $o(1) \to 0$ in probability as $n \to \infty$. From the standard argument [7]–[8], we need only to show that the middle term on the right hand side of (4.5) converge in distribution to a normal random variable with 0 mean and appropriate variance. Since $\sum_{j=1}^{N_i(n)} \Phi_{N_i(n)|j}^i \frac{\alpha_{\tau_j^i}^i}{\sqrt{\tau_j^i}} \xi_{\tau_j^i}^i$ and $\sum_{j=1}^{N_i(n)} \Phi_{N_i(n)|j}^i \frac{1}{\sqrt{\tau_j^i}} \xi_{\tau_j^i}^i$ converge in distribution to the same limit, we need only to prove the claim for the latter one. In order to do so, we first establish

Lemma 4. *If* (A3)– (A6) *hold, then* $\sum_{j=1}^{n} \Phi_{n|j}^i \frac{1}{\sqrt{\tau_j^i}} \xi_{\tau_j^i}^i$ *converges in distribution to a random variable which is normally distributed with* 0 *mean and variance* s_i*:*

$$s_i = \int_0^\infty e^{a_i t} \Sigma_i e^{a_i t} dt$$

where $a_i = \frac{\mu^{i\prime}}{2} - \lambda_i$, $\Sigma_i = \sum_{j,l=1}^{L} c_j r^i c_l$.

The essential idea of the proof can be found in [7]. We will not dwell on it here.

Lemma 5. *Under the same assumptions of Lemma* 4,

$$\sum_{j=1}^{N_i(n)} \Phi_{N_i(n)|j}^i \frac{1}{\sqrt{\tau_j^i}} \xi_{\tau_j^i}^i$$

converges in distribution to a normal random variable with 0 *mean and variance* $\frac{s_i}{\mu_i}$.

The proof for this lemma can be found in [9] Theorem 17.1. Again we omit the detail.

The independence of ξ^i and ξ^j implies the independence of the limits $\tilde{\Phi}_n^i$ and $\tilde{\Phi}_n^j$ for $i \neq j$, where $\tilde{\Phi}_n^i$ is defined as

$$\tilde{\Phi}_n^i = \sum_{j=1}^{N_i(n)} \Phi_{N_i(n)|j}^i \frac{1}{\sqrt{\tau_j^i}} \xi_{\tau_j^i}^i.$$

Combine this with Lemma 3 and Lemma 4, we have

Theorem 2. *If* (A3)–(A6) *hold, then* $\sqrt{n} x_n \xrightarrow{D} N(0, \Sigma)$, *where* $\Sigma = \operatorname{diag}(\frac{s_1}{\mu_1}, \cdots, \frac{s_r}{\mu_r})$.

Now we are in a position to eliminate the assumption (A6). As we commented before, if D is a symmetric negative definite matrix, then equation (3.5) holds. To obtain asymptotic normality, we need to examine the following term:

$$\sum_{j=1}^{N_i(n)} \Phi_{N_i(n)|j}^i \frac{1}{\sqrt{\tau_j^i}} \sum_{k=1}^{r} P_{i,k}^{-1} \xi_{\tau_j^i}^k$$

where $P^{-1}_{i,k}$ is the (i,k)th entry of the matrix P^{-1}. Similar kind of analysis as above gives us:

Theorem 3. *Under assumptions* (A3)–(A5),

$$\sqrt{n}P^{-1}x_n \xrightarrow{D} N(0,\tilde{\Sigma})$$

where $\tilde{\Sigma}$ is the convariance matrix, and the ijth entry is given by

$$\tilde{\Sigma}_{ij} = \frac{1}{\sqrt{\mu^i \mu^j}} \sum_{k=1}^{r} P^{-1}_{ik} P^{-1}_{jk} s_k.$$

By virtue of the well-known Slustky's lemma (cf. [10]), the following corollary holds.

Corollary. $\sqrt{n}x_n \xrightarrow{D} N(0, P\tilde{\Sigma}P')$.

The asymptotic normality can also be established via the method of weak convergence. The interested readers are referred to [5].

References

[1] Yin Gang and Zhu Yun-min, On w.p.1 convergence of a parallel stochastic approximation algorithm, LCDS#87-17, Brown Univ., Providence, RI., 1987.

[2] J. N. Tsitsiklis, Problems in decentralized decision making and computation, Ph. D. thesis, Elect. Eng. Dept., M.I.T., Cambridge, MA, 1984.

[3] D. Bertsekas, J. N. Tsitsiklis and M. Athans, Convergence theories of distributed iterative processes; a survey, Technical Report for Information and Decision Systems, M.I.T., Cambridge, MA, 1984.

[4] H. J. Kushner and Yin Gang, Asymptotic properties of distributed and communicating stochastic approximation algorithms, *SIAM J. on Control and Optimization*, 25 (1987), 1266–1290.

[5] H. J. Kushner and Yin Gang, Stochastic approximation algorithms for parallel and distributed processing, *Stochastics*, 22 (1987), 219–250.

[6] Zhu Yun-min, Extensions of the relation of a series to infinite products and the convergence of a class of recursive algorithms, *J. Numer Math.*, 7 : 4 (1985), 369–376.

[7] H. F. Chen and Zhu Yun-min, Asymptotic properties of a stochastic approximation procedure with randomly varying truncations, *Acta Mathematica Scientia*, 7 (1987), 4.

[8] M. B. Nevelson and R. Z. Hasminskii, Stochastic Approximation and Recursive Estimations, *Translation of Math. Monographs*, 47, AMS, 1976.

[9] P. Billingsley, Convergence of Probability Measures, John Wiley, 1968.

[10] Y. S. Chow and H. Teicher, Probability Theory, Springer–Verlag, 1978.