Journal of Computational Mathematics, Vol.12, No.1, 1994, 21–30.

# STOCHASTIC APPROXIMATION IN REAL TIME: A PIPE LINE APPROACH*

Zhu Yun-min[1)]
(*Inst. of Math. Scis., Chengdu Branch, Academia Sinica, Chengdu, Sichuan, China*)

Yin Gang[2)]
(*Department of Mathematics, Wayne State University, Detroit, USA.*)

**Abstract**

A new approach for stochastic approximation in real time is developed. A number of processors are simultaneously active to carry out a computing task. All processors work on the same system with different starting time. After each iteration, computed data are passed to the next processor on line. Interacting tasks and iterative instructions are carried through pipelining of computation and communication. Asymptotic properties of the algorithm are developed, and comparisons of the performance between the new algorithm and the classical one are made.

## 1. Introduction

The objective of this work is to study stochastic approximation in real time. A pipe line approach is suggested. Asymptotic properties of the procedure are developed, and comparisons of rate of convergence with the classical algorithms are made.

Let $x \in \mathbf{R}^L$, and $f(\cdot) : \mathbf{R}^L \mapsto \mathbf{R}^L$. The traditional stochastic approximation methods deal with the problem of finding the roots of $f(x) = 0$ by using noisy measurements $Y_{\hat{n}}$ via the recursive procedure

$$X_{\hat{n}+1} = X_{\hat{n}} + a_{\hat{n}} Y_{\hat{n}}, \ Y_{\hat{n}} = f(X_{\hat{n}}) + \xi_{\hat{n}} \tag{1.1}$$

where the gains $a_{\hat{n}}$ satisfy $a_{\hat{n}} > 0$, $\sum a_{\hat{n}} = \infty$, $\frac{a_{\hat{n}} - a_{\hat{n}+1}}{a_{\hat{n}}} \xrightarrow{\hat{n}} 0$, and $\xi_{\hat{n}}$ represent the measurement errors. Various successful applications of the stochastic approximation methods have been reported[1,2].

The Robbins-Monro (RM) algorithm (1.1) can be thought of as a two phase operation. For each iteration, the first phase is to take measurement $Y_{\hat{n}}$, and the next step is to form the new estimate $X_{\hat{n}+1}$ by means of addition. Usually, most of the computation time is spent on the process of collecting data $Y_{\hat{n}}$. The second step $X_{\hat{n}} + a_{\hat{n}} Y_{\hat{n}}$ is less

* Received May 21, 1991.

[1)] The Project Supported in part by Science Foundation of Academia Sinica.

[2)] The Project Supported by National Natural Science Foundation of China, and by Wayne State University.

time consuming. This feature is, however, not reflected in (1.1) since $\hat{n}$ represents the iteration number rather than real time.

In recent years, parallel processing methods have drawn much of attention. Starting from [3], several stochastic approximation algorithms have been developed for parallel and distributed computing[4–6]. An extensive survey can be found in [7]. In view of the recent developments, and motivated by the idea of pipelining of computation for large scale parallelization[8], a new algorithm for stochastic approximation in real time is suggested here. In lieu of using a single processor alone as in the traditional setting, a number of identical processors are utilized to update the same system. The processors are lined up as on a production line. After one step iteration is completed, the newly computed data are passed to the next processor on the line. Each processor, repeatedly, executes the same instruction on successive observed data and data received from the preceding processor.

Assume that a single processor needs $r$ units of time for a phase one operation and 1 unit of time for an addition; let $n$ denote real time. Instead of algorithm (1.1), $r+1$ processors will be used to carry out the computing task. All $r+1$ processors work on the same system vector $X$, and communicate with each other through pipe line structure. In spite of different time indices, the scheme is the same for all the processors. Thus, it suffices to use a single formula to describe the procedure. The initial conditions are given for $X_1, X_2, \cdot, X_{r+2}$. The observation at time $n$ is $Y_n$. The algorithm is given by

$$X_{n+1} = X_n + a_{n-r} Y_{n-r} \tag{1.2}$$

where the gains $a_n$ are as before with $\hat{n}$ replaced by $n$. When $r = 0$, the above algorithm formally reduces to the classical stochastic approximation procedure.

In the proposed algorithm, the overall system consists of a number of parallel processors connected through communications. The length of a computation cycle is equal to the number of processors participating in the computation, which is also equal to the time required for a single processor to complete one iteration. The notation $\{Y_{n-r}\}$ means that the measurement was begun $r$ (time) units before, and completed at time $n$. We shall call such measurements "delayed" measurements. In fact, in various situations, the observed signals are rarely available immediately without any delays. For example, for the Viterbi decoding algorithm, the desired signals are not available until several symbol intervals later. Therefore, even for a single processor alone, an algorithm with delay seems to be a more natural model.

The remainder of this paper is arranged as follows. A modified algorithm is given first, and the strong convergence is obtained. An order of magnitude estimate of the algorithm is derived in Section 3, and asymptotic normality is proved in Section 4. Further discussions and comparisons of rates of convergence are given in Section 5. Finally, an appendix is included. In the sequel, "′" stands for the transpose of a matrix; $g_x(\cdot)$ denotes the gradient of a function $g(\cdot)$ and $K$ denotes a generic positive constant with possibly different values.

## 2. A Modified Algorithm

Let $f(\cdot) : \mathbb{R}^L \mapsto \mathbb{R}^L$ be a continuous function. Denote the set of zeros of $f$ by $Z = \{x \in \mathbb{R}^L; f(x) = 0\}$. The following conditions are needed in the subsequent development.

(A1) $\xi_n = \alpha_n + \beta_n$ such that $\sum_n a_n \alpha_n$ converges a.s. and $\beta_n \xrightarrow{n} 0$ a.s.

(A2) There is a twice continuously differentiable Liapunov function $V(\cdot) : \mathbb{R}^L \mapsto \mathbb{R}$, such that (1) $V(x) \geq 0$, $\forall x$, and $V(x) \to \infty$ as $|x| \to \infty$; (2) For some $x^*$, there exists a $M > 0$, with $|x^*| \leq M$. $V(x^*) \neq \inf_{|x|=M} V(x)$. Moreover, there exists a $\delta_1 > 0$, such that

$$\delta_1 \in \Big(V(x^*), \inf_{|x|=M}\Big) \cup \Big(\inf_{|x|=M} V(x), V(x^*)\Big) - \{V(x); x \in Z\}.$$

(3) $V'(x)f(x) < 0$, for all $x \notin Z$.

The consistency of stochastic approximation algorithms without assuming a priori boundedness on $f(\cdot)$ was proved in [9]. We adapt the truncation method in our paper to prove the convergence for the real time stochastic approximation algorithm.

Let $\{M_n\}$ be a sequence of increasing positive real numbers, such that $M_n \xrightarrow{n} \infty$. Let $\{\sigma(n)\}$ be a sequence of random variables given by

$$\sigma(0) = 0, \quad \sigma(n+1) = \sigma(n) + I_{\{|X_n + a_{n-r}Y_{n-r}| > M_{\sigma(n)}\}}, \tag{2.1}$$

where $I_A$ denotes the indicator function of a set $A$. (2.1) provides us with a sequence of truncation bounds. Now, define the algorithm with randomly varying truncations as

$$X_{n+1} = (X_n + a_{n-r}Y_{n-r}) I_{\{|X_n + a_{n-r}Y_{n-r}| \leq M_{\sigma(n)}\}} + x^* I_{\{|X_n + a_{n-r}Y_{n-r}| > M_{\sigma(n)}\}}. \tag{2.2}$$

**Theorem 2.1.** *Under* (A1) *and* (A2), *for any initial conditions, $X_j$, $1 \leq j \leq r+2$, $\lim_{n\to\infty} d(X_n, Z) = 0$ a.s. where $d(\cdot,\cdot)$ denotes the distance function.*

To prove the theorem, we first state a lemma, which is a modification of Lemma 1 and Lemma 2 in [9]. The detailed proofs are omitted.

**Lemma 2.2.** *Let $\{X_{n_k}\}$ be a convergent subsequence of $\{X_n\}$, and let $X_{n_k - j}$, $j \leq r$ be bounded uniformly in $k$. Suppose that the conditions of Theorem* 2.1 *are satisfied. Then there exist $c_1, c_2 > 0$, $\Delta > 0$; and for any $\eta$ with $0 \leq \eta < \Delta$, there exists a $K_\eta$, such that for any $k \geq K_\eta$, and for $m(n,\eta) = \max\{m;\ \sum_{j=n}^{m} a_j \leq \eta\}$,*

$$\Big| \sum_{j=n_k}^{m} a_{j-r} Y_{j-r} \Big| \leq c_1, \ m \in [n_k - r, m(n_k, \eta)], \tag{2.3}$$

$$|X_m - X_{n_k}| \leq c_2 \eta, \ m \in [n_k - r, m(n_k, \eta) + 1]. \tag{2.4}$$

It is easily seen that $\sigma(n)$ is a non-decreasing sequence. Thus, either $\sigma(n) \xrightarrow{n} \sigma$ a finite limit a.s., or it grows without bound. We demonstrate that the first alternative holds.

**Lemma 2.3.** *Under the conditions of Theorem* 2.1, *$\sigma < \infty$ a.s.*

*Proof.* Without loss of generality, we may assume that $V(x^*) < \inf_{|x|=M} V(x)$. As a consequence, there exists $[\delta_1, \delta_2] \subset [V(x^*), \inf_{|x|=M} V(x)]$ and $\delta_1 > V(x^*), \delta_2 <$

$\inf_{|x|=M} V(x), \delta_1 \notin V(E)$, for, if Lemma 2.3 were not true, then starting from $x^*$, $X_n$ would be across the sphere $\{x; |x| = M\}$ infinitely often, and hence $V(X_n)$ would be across the interval $[\delta_1, \delta_2]$ infinitely often from the left. Thus, there exist subsequences $\{X_{l_k}\}$, $\{X_{n_k}\}$ and $\{X_{m_k}\}$ of $\{X_n\}$, such that

$$l_k < n_k < m_k,\ X_{l_k} = x^*,\ X_i \le M,\ \forall k,\ l_k \le i \le n_k - 1, \tag{2.5}$$

$$V(X_{n_k-1}) < \delta_1 \le V(X_i) \le \delta_2 \le V(X_{m_k}),\ \forall k,\ n_k \le i \le m_k - 1. \tag{2.6}$$

From (2.2) and (2.5), it is clear that

$$X_{i+1} - X_i \xrightarrow{k} 0, \quad i \in [l_k, n_k]. \tag{2.7}$$

The continuity of $V(\cdot)$ then implies that $V(X_{n_k}) \xrightarrow{k} \delta_1$. Since $|X_{n_k}| \le M$, we can extract another convergent subsequence with limit $\bar{x}$. For notational simplicity, we still denote this sequence by $\{X_{n_k}\}$. By virtue of (2.6), $V(\bar{x}) = \delta_1$. Owing to (2.8), $\{X_{n_k}\}$ is a subsequence which satisfies the assumption of Lemma 2.1. From here on, we can use the same arguments as in [9]. Some details are omitted.

It follows from the above lemma that there is a $\kappa_1 \ge 1$, such that for all $n \ge \kappa_1$,

$$|X_n| \le M_\sigma \text{ and } X_{n+1} = X_n + a_{n-r} Y_{n-r}. \tag{2.8}$$

Thus, eventually the truncations will be terminated, and the algorithm is bounded uniformly for large $n$ with probability one.

*Proof of Theorem* 2.1. Owing to (2.8) and (A1), and the fact that only finitely many terms are involved between time $n - 2r$ and $n - r$, for some $\kappa_2 \ge \kappa_1 + 2r$ and all $n \ge \kappa_2$,

$$X_n - X_{n-r} = \sum_{j=n-2r}^{n-r} a_j f(X_j) + \sum_{j=n-2r}^{n-r} a_j \alpha_j + \sum_{j=n-2r}^{n-r} a_j \beta_j \xrightarrow{n} 0 \text{ a.s.} \tag{2.9}$$

The continuity of $f(\cdot)$ then yields $f(X_n) - f(X_{n-r}) \xrightarrow{n} 0$ a.s. Next, for $n \ge \kappa_2$, rewrite (2.8) as

$$X_{n+1} = X_n + a_n f(X_n) + a_n \tilde{\xi}_n \tag{2.10}$$

where

$$\tilde{\xi}_n = \tilde{\alpha}_n + \tilde{\beta}_n \text{ with } \tilde{\alpha}_n = \frac{a_{n-r}}{a_n} \alpha_{n-r},$$

$$\tilde{\beta}_n = \frac{a_{n-r} - a_n}{a_n} f(X_{n-r}) + (f(X_{n-r}) - f(X_n)) + \frac{a_{n-r}}{a_n} \beta_{n-r}.$$

The choice of the gain $\{a_n\}$ implies that $(a_{n-r} - a_n)/a_n \xrightarrow{n} 0$. Therefore, the boundedness of $X_{n-r}$ and the choice of $a_n$ imply that the first term in $\tilde{\beta}_n$ tends to 0 a.s. The argument following (2.9) then implies that the second term also tends to zero. Noticing that $a_{n-r}/a_n$ is bounded, (A1) then yields that the third term goes to 0. As a consequence, $\tilde{\beta}_n \xrightarrow{n} 0$ a.s. It is readily seen that $\sum_j a_j \tilde{\alpha}_j$ converges a.s. Owing to (2.10), for all $n \ge \kappa_2$, the algorithm can be viewed as a standard stochastic approximation algorithm with measurement noise $\{\tilde{\xi}_n\}$. The technique of the ODE approach (cf. [1, 2]) is in force. By using this approach and similarly as in [9], we obtain $d(X_n, Z) \xrightarrow{n} 0$ a.s. as desired.

## 3. An Order Estimate

For simplicity, we shall assume $a_n = \frac{1}{n}$ henceforth. A similar result can be obtained for $a_n = \frac{C}{n^{\kappa}}$, $0 < \kappa \leq 1$, and $C$ a positive constant. It is well-known, for the classical stochastic approximation algorithm with a single processor and with $r = 0$, that $\sqrt{\hat{n}}(X_{\hat{n}+1} - \theta)$ converges in distribution to a normal random variable[1,10,11]. We show that a similar result still holds for the "real time" algorithm considered in this work. The asymptotic covariance matrix is a standard measure of rate of convergence and can be used as a basis for comparison of different algorithms.

In what follows, the rate of convergence is studied in two steps. In this section, we derive an order of magnitude estimate which is important in the subsequent development. Then, in Section 4, the asymptotic normality is established. In addition to (A1) and (A2), assume

(A3) $Z = \{\theta\}$ and $f(x) = F(x - \theta) + \delta(x)$ where $\delta(x) = O(|x - \theta|^{1+\gamma})$, for some $\gamma > 0$. Moreover, all eigenvalues of $A = \frac{I}{2} + F$ have negative real parts.

(A4) Let $E^{\mathcal{F}_n}$ denote the conditioning on the $\sigma$-algebra $\mathcal{F}_n$ generated by past data up to time $n$. The following inequalities hold uniformly in $n$ :

$$E\sum_{k=n}^{\infty} \frac{1}{j}|E^{\mathcal{F}_n}\xi_{k-r}| = O\,(1/n)\,, \quad E\sum_{k=n}^{\infty} \frac{1}{j}|E^{\mathcal{F}_n}\xi_{n-r}\xi_{k-r}| = O\,(1/n)\,.$$

**Remark.** If the noise processes satisfy

$$E\sum_{k=n}^{\infty} |E^{\mathcal{F}_n}\xi_{k-r}| < \infty \text{ and } E\sum_{k=n}^{\infty} |E^{\mathcal{F}_n}\xi_{n-r}\xi_{k-r}| < \infty \text{ uniformly in } n,$$

then (A4) holds. In the next section, we shall consider $\varphi$-mixing processes. For such processes, with suitable conditions on the mixing measures, the above inequalities hold.

**Theorem 3.1.** *Let* (A1)–(A4) *be satisfied, and suppose that there exists a* $\lambda > 1$ *such that* $V_x'(x)f(x) < -\lambda V(x)$. *Then*

$$EV(X_{n+1}) = O\,(1/n)\,, \textit{ for sufficiently large } n. \tag{3.1}$$

**Corollary 3.2.** If there is a positive definite matrix $R$, such that $V(x) = (x - \theta)'R(x - \theta) + o(|x - \theta|^2)$, then

$$E|X_{n+1} - \theta|^2 = O\,(1/n) \text{ for sufficiently large } n. \tag{3.2}$$

*Proof.* We prove (3.1) only. This is a modification of the result obtained in [11]. By virtue of Lemma 2.3, there is a $\kappa_1$, such that $\{X_n\}$ and $\{V(X_n)\}$ are both bounded for all $n \geq \kappa_1$. Owing to (2.8), for $n \geq \kappa_2 \geq \kappa_1 + 2r$,

$$E^{\mathcal{F}_n}V(X_{n+1}) - V(X_n) = \frac{1}{n}V_x'(X_n)f(X_n) + \frac{1}{n-r}V_x'(X_n)\xi_{n-r} + O\left(\frac{1}{n^2}\right)$$
$$+ V_x'(X_n)\left(\frac{1}{n-r}f(X_{n-r}) - \frac{1}{n}f(X_n)\right). \tag{3.3}$$

Define
$$V_1(n) = \sum_{k=n}^{\infty} \frac{1}{k-r} E^{\mathcal{F}_n} V_x'(X_n)\xi_{k-r}. \tag{3.4}$$

By virtue of (A4),
$$E|V_1(n)| \le \frac{K}{n}(1 + V(X_n)). \tag{3.5}$$

Define $\tilde{V}(n) = V(X_n) + V_1(n)$. Then
$$E^{\mathcal{F}_n}\tilde{V}(n+1) - \tilde{V}(n) = \frac{1}{n}V_x'(X_n)f(X_n) + V_x'(X_n)\left(\frac{1}{k-r}f(X_{k-r}) - \frac{1}{n}f(X_n)\right)$$
$$+ O(1/n^2) + K\sum_{k=n+1}^{\infty}\frac{1}{n-r}\left(\frac{1}{k-r}|E^{\mathcal{F}_n}\xi_{k-r}\xi_{n-r}'| + \frac{1}{k-r}|E^{\mathcal{F}_n}\xi_{k-r}|\right). \tag{3.6}$$

Notice that
$$X_{n-r} - X_n = -\left(\sum_{k=n-2r}^{n-r-1}\frac{1}{k}F(X_k - \theta) + \sum_{k=n-2r}^{n-r-1}\frac{1}{k}\xi_k + \sum_{k=n-2r}^{n-r-1}\frac{1}{k}\delta(X_k)\right). \tag{3.7}$$

In view of (A3), (A4) and (3.7),
$$E\left|V_x'(X_n)\left(\frac{1}{k-r}f(X_{n-r}) - \frac{1}{n}f(X_n)\right)\right| = O(1/n^2)(1 + EV(X_n)). \tag{3.8}$$

By virtue of (3.5), (3.7), (3.8) and (A4), taking expectation in (3.6) yields
$$E\tilde{V}(n+1) - E\tilde{V}(n) \le -\lambda E\tilde{V}(n) + O(1/n^2)(1 + E\tilde{V}(n)). \tag{3.9}$$

(3.9) in turn yields that for some $\kappa_3 \ge 1$, $\lambda_1 > 1$, and all $n \ge \kappa_3$,
$$E\tilde{V}(n+1) \le (1 - \lambda_1/n)E\tilde{V}(n) + O\left(1/n^2\right).$$

Choose $\kappa_4 = \max\{\kappa_2, \kappa_3\}$. Then
$$E\tilde{V}(n+1) \le \prod_{k=\kappa_4+1}^{n}(1 - \lambda_1/k)E\tilde{V}(\kappa_4) + K\sum_{k=\kappa_4}^{n}\frac{1}{k^2}\prod_{j=k+1}^{n}(1 - \lambda_1/j). \tag{3.10}$$

Using a familiar inequality $\prod_{j=1}^{n}(1 - \lambda_1/j) \le \frac{K}{n^{\lambda_1}}$, we obtain $E\tilde{V}(n+1) \le O\left(n^{-\lambda_1}\right) + O\left(n^{-1}\right) = O\left(n^{-1}\right)$. Finally, in view of (3.5), Eq. (3.1) holds for $n$ large enough. The proof is concluded.

## 4. Asymptotic Distribution

We show that $\sqrt{n}(X_{n+1} - \theta) \sim N(0, \Sigma)$ in this section. It is assumed that equation (3.2) holds throughout the rest of the paper. In addition, assume

(A5) $\{\xi_n\}$ is a stationary stochastic process with $E\xi_1 = 0$, such that for some $a > 0$, $\sup_n E|\xi_n|^{2+a} < \infty$ a.s. Define $\mathcal{F}_n = \sigma\{\xi_k; k \leq n\}$, $\mathcal{F}^n = \sigma\{\xi_k; k \geq n\}$. For $j \geq 1$, let $\varphi(j)$ be given by $\varphi(j) = \sup_{A \in \mathcal{F}^{j+n}} E^{\frac{1+a}{2+a}}[P(A|\mathcal{F}_n) - P(A)]^{\frac{2+a}{1+a}}$ and suppose $\sum_j [\varphi(j)]^{\frac{a}{1+a}} < \infty$.

We claim that the following lemma holds. The proof is provided in the appendix.

**Lemma 4.1.** *Suppose that* (A1)-(A3), (3.2) *are satisfied, and suppose that* (A5) *holds. Then there is an* $m \geq \kappa_4$, *such that for all* $n - r \geq m$,

$$\sqrt{n}(X_{n+1} - \theta) = \sqrt{n-r} \sum_{k=m}^{n-r} \frac{1}{k} A_{n-r,k} \xi_k + o(1) \tag{4.1}$$

*where* $o(1) \xrightarrow{n} 0$ *in probability and*

$$A_{jk} = \begin{cases} \prod_{l=k+1}^{j} (I + F/l); & \mathit{ifj} > k \\ I; & \mathit{ifj} = k. \end{cases}$$

**Theorem 4.2.** *Under the conditions of Lemma* 4.1, $\sqrt{n}(X_{n+1} - \theta) \sim N(0, \Sigma)$, *where*

$$\Sigma = \int_0^\infty e^{At} \bar{S} e^{A't} dt \ \textit{with}\ \bar{S} = E(\xi_1 \xi_1') + \sum_{k=2}^{\infty} E(\xi_1 \xi_k') + \sum_{k=2}^{\infty} E(\xi_k \xi_1'). \tag{4.2}$$

In lieu of examining (4.1) directly, we consider the interpolated process

$$W_n(t) = \frac{[nt]}{\sqrt{n}} \sum_{k=1}^{[nt]} \frac{1}{k} A_{[nt]k} \xi_k \ \text{ for } t \in [0,1] \tag{4.3}$$

where $[z]$ denotes the largest integer which is smaller than or equal to $z$. We shall first show that $W_n(\cdot)$ converges weakly to $W(\cdot)$. The proof is an application of the functional central limit theory approach and the methods of weak convergence in [12].

**Lemma 4.3.** *Let* $B_n(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{[nt]} \xi_j$. *Under* (A5), $B_n(\cdot)$ *converges weakly to* $B(\cdot)$, *a Brownian motion with covariance* $t\bar{S}$.

Using summation by parts in (4.3),

$$W_n(t) = B_n(t) + (I + F)[nt] \sum_{k=1}^{[nt]-1} \frac{1}{k(k+1)} A_{[nt](k+1)} B_n(k/n). \tag{4.4}$$

**Theorem 4.4.** *Under* (A1)-(A3), (3.2) *and* (A5), $W_n(\cdot)$ *converges weakly to* $W(\cdot)$, *a Gauss-Markov process with* $W(t) = \int_0^t e^{-(I+F)(\ln u - \ln t)} dB(u)$ *and* $B(\cdot)$ *given by Lemma* 4.3.

*Proof.* The weak limit of the second term on the right of (4.4) is the same as that of

$$(I+F)[nt]\sum_{k=1}^{[nt]-1}\frac{1}{k(k+1)}e^{-F\left(\ln\frac{k+1}{[nt]}\right)}B\left(k/n\right)$$
$$\to (I+F)\int_0^1 \frac{1}{u^2}e^{-F\ln u}B(ut)du. \tag{4.5}$$

By virtue of Lemma 4.5, integration by parts and change of variables yield

$$W(t)=\int_0^1 e^{-(I+F)\ln u}dB(ut)=\int_0^t e^{-(I+F)(\ln u-\ln t)}dB(u). \tag{4.6}$$

Theorem 4.4 thus follows.

## 5. Further Discussions

The rate of convergence is determined by the largest number $\rho$, with $0<\rho\leq 1$ for which the asymptotic part of $n^{\rho}(X_{n+1}-\theta)$ converges to a non-degenerate and "stable" process (cf. [1] Chapter VII); with the same scaling factor, the asymptotic covariance matrices can be used for comparisons of rates of convergence.

Since $r+1$ processors are used in the new algorithm and arranged in pipe line, the iteration time for each step is one unit only as illustrated in the introduction section. Setting $t=1$ in $W_n(t)$ and in view of Lemma 4.1,

$$\sqrt{n}(X_{n+1}-\theta)\sim N(0,\Sigma), \tag{5.1}$$

with $\Sigma=E(W(1)W'(1))=\int_0^\infty e^{Au}\bar{S}A^{A'u}du$.

For the classical RM algorithm, let $\hat{n}$ be the iteration number. We have

$$\sqrt{\hat{n}}(X_{\hat{n}+1}-\theta)\sim N(0,\Sigma). \tag{5.2}$$

Since the actual time $n$ for $\hat{n}$ iterations is $n=(r+1)\hat{n}$, a fair comparison should involve comparing (5.1) with $\sqrt{(r+1)\hat{n}}(X_{\hat{n}+1}-\theta)\sim N(0,\Sigma')$. Due to (5.2) and the well-known Slutsky's theorem, $\Sigma'=(r+1)\Sigma$. This indicates that the asymptotic covariance of the traditional RM algorithm is $r+1$ times as large as the algorithm proposed in this work. As a result, the new algorithm provides a speedup with a factor $r+1$. Therefore, the new algorithm can be thought of as an acceleration procedure. If the classical procedure is used with a single processor, then a certain amount of time is spent on the waiting for the required data to become available. In the newly developed algorithm, at any given time, there is always one processor doing the phase 2 computation (addition), and all the others are in phase 1—the data collection mode. In this way, the amount of real time required for iteration is reduced from $r+1$ units to 1 unit only. When one processor communicates its partial result to another processor, all other processors, including the one being communicated, are still in operation. This fact puts the communication penalty in a relatively insignificant level. As a result, the communication penalty will not prevent the concurrent utilization of a large number of processors in parallel when the underlying computing task is large.

We hope that the idea exploited in this paper will open up a new domain in studying real time stochastic approximation problems. Such study is very important for various applications.

## Appendix

*Proof of Lemma* 4.1. For $n - r \geq m$,

$$\sqrt{n}(X_{n+1} - \theta) = \sqrt{n}A_{n-r,m-1}(X_{m-1} - \theta) + \sqrt{n}\sum_{k=m}^{n-r}\frac{1}{k}A_{n-r,k}\delta(X_{k+r})$$
$$+ \sqrt{n}\sum_{k=m}^{n-r}\frac{1}{k}A_{n-r,k}\Big(f(X_k) - f(X_{k+r})\Big) + \sqrt{n}\sum_{k=m}^{n-r}\frac{1}{k}A_{n-r,k}\xi_k. \tag{a.1}$$

By virtue of (3.2), the first two terms on the right-hand side of ($a$.1) tend to 0 in probability. As for the third term,

$$\sqrt{n}\sum_{k=m}^{n-r}\frac{1}{k}A_{n-r,k}\left(f(X_k) - f(X_{k+r})\right) = \sqrt{n}\sum_{k=m}^{n-r}\frac{1}{k}A_{n-r,k}F(X_k - X_{k+r})$$
$$+ \sqrt{n}\sum_{k=m}^{n-r}\frac{1}{k}A_{n-r,k}\delta(X_k) - \sqrt{n}\sum_{k=m}^{n-r}\frac{1}{k}A_{n-r,k}\delta(X_{k+r}). \tag{a.2}$$

It can be shown that the last two terms on the right-hand side of (a.2) tend to 0 in probability.

Next, we examine the first term on the right-hand side of ($a$.2). In view of (3.7),

$$\sqrt{n}\sum_{k=m}^{n-r}\frac{1}{k}A_{n-r,k}F(X_k - X_{k+r}) = -\sqrt{n}\sum_{k=m}^{n-r}\frac{1}{k}A_{n-r,k}F\sum_{j=k-r}^{k-1}\frac{1}{j}F(X_j - \theta)$$
$$+ \sum_{k=m}^{n-r}\frac{1}{k}A_{n-r,k}F\sum_{j=k-r}^{k-1}\frac{1}{j}\xi_j + \sum_{k=m}^{n-r}\frac{1}{k}A_{n-r,k}F\sum_{j=k-r}^{k-1}\frac{1}{j}\delta(X_j). \tag{a.3}$$

The definition of $A_{n-r,k}$ implies that $|A_{n-r,k}| \leq \left(\frac{k}{n-r}\right)^{\alpha_1}$ for some $\alpha_1 > 0$. Consequently, for any $\eta > 0$, by virtue of equation (3.2) and Chebyshev's inequality,

$$P\Big\{\Big|\sqrt{n}\sum_{k=m}^{n-r}\frac{1}{k}A_{n-r,k}\sum_{j=k-r}^{k-1}\frac{F^2}{j}(X_j - \theta)\Big| > \eta\Big\}$$

$$\leq \sqrt{\frac{n}{n-r}}\frac{K}{\eta(n-r)^2}\sum_{k=m}^{n-r}\left(\frac{k}{n-r}\right)^{\alpha_1 - \frac{5}{2}} \xrightarrow{n} 0.$$

In addition, the other two terms on the right-hand side of ($a$.3) tend to 0 in probability. Furthermore, using similar arguments, we can show

$$\sqrt{n}\sum_{k=m}^{n-r}\frac{1}{k}A_{n-r,k}\xi_k = \sqrt{n-r}\sum_{k=m}^{n-r}\frac{1}{k}A_{n-r,k}\xi_k + o(1) \tag{a.4}$$

with $o(1)\xrightarrow{n}0$ in probability. The proof of the lemma is thus concluded.

Acknowledgement. The authors wish to thank Professor Chen Han-Fu for several discussions on this work and for some valuable suggestions.

## References

[1] H.J. Kushner and D.S. Clark, Stochastic Approximation for Constrained and Unconstrained Systems, Springer-Verlag, Berlin, 1978.

[2] L. Ljung, Analysis of recursive stochastic algorithms, *IEEE Trans. on Automat. Control*, AC-22 (1977), 551–575.

[3] J.N. Tsitsiklis, D.P. Bertsekas and M. Athans, Distributed asynchronous deterministic and stochastic gradient optimization algorithms, *IEEE Trans. on Automat. Control*, AC-31 (1986), 803–812.

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

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

[6] G. Yin and Y.M. Zhu, On w.p.1 convergence of a parallel stochastic approximation algorithm, *Probab. Eng. Inform. Sci.*, **3** (1989), 55–75.

[7] G. Yin, Recent progress in parallel stochastic approximations, Topics in Stochastic Systems: Modelling, Estimation and Adaptive Control, L. Gerencsér and P.E. Caines Eds., Springer-Verlag, 1991, 159–184.

[8] D.P. Bertsekas and J.N. Tsitsiklis, Parallel and Distributed Computing, Prentice-Hall, New Jersey, 1989.

[9] H.F. Chen and Y.M. Zhu, Stochastic approximation procedures with randomly varying truncations, *Scientia Sinica* (series A), **29** (1986), 914–926.

[10] H.J. Kushner, Approximation and Weak Convergence Methods for Random Processes with Applications to Stochastic Systems Theory, MIT Press, Cambridge, 1984.

[11] H.J. Kushner and H. Huang, Asymptotic properties of stochastic approximations with constant coefficients, *SIAM J. Control Optim.*, **19** (1981), 86–105.

[12] S. Ethier and T.G. Kurtz, Markov Processes: Characterization and Convergence, Wiley & Sons, New York, 1986.