15th Triennial World Congress, Barcelona, Spain

STOCHASTIC APPROXIMATION ALGORITHMS WITH EXPANDING TRUNCATIONS[1]

Han-Fu Chen *

* *Laboratory of Systems and Control*
Institute of Systems Science
Academy of Mathematics and Systems Science
Chinese Academy of Sciences

Abstract: The purpose of stochastic approximation is to find the roots of an unknown function $f(\cdot)$, which can be observed, but the observations are corrupted by errors. General convergence theorems for stochastic approximation algorithms with expanding truncations are presented. The observation errors are allowed to include both random noise and structural uncertainties. The conditions imposed on the observation errors are the weakest possible, while the function $f(\cdot)$ is only required to be measurable and locally bounded. Applications of the general convergence theorems to optimization and signal processing demonstrate the strong points of results given in the paper.

Keywords: Stochastic approximation, expanding truncation, convergence, optimization, adaptive filtering, channel identification

1. INTRODUCTION

Stochastic approximation (SA) is devoted to finding zeros of an unknown function $f(\cdot)$ which can be observed, but the observations are corrupted by errors. The Robbins-Monro (RM) algorithm [1] provides the following estimate for the root of $f(\cdot)$:

$$x_{k+1} = x_k + a_k y_{k+1} \tag{1}$$

where a_k is the step size, and y_{k+1} is the observation at time $k+1$, which is given by

$$y_{k+1} = f(x_k) + \epsilon_{k+1} \tag{2}$$

where ϵ_{k+1} is the observation noise.

Until mid-seventies of twentieth century the probabilistic method was the main approach for convergence analysis of SA algorithms. By this method, the linear growth rate of $f(\cdot)$ and independence or martingale property for the observation noise are assumed, in addition to the usual condition concerning existence of a Lyapunov function, which normally has to be assumed in all analysis methods for SA algorithms [2]. This type of works was well summarized in [3].

Starting from mid-seventies the ordinary differential equation (ODE) method was developed for convergence analysis of SA algorithms [4,5] by observing that the tail part of the interpolating function of $\{x_k\}$ with step size $\{a_k\}$ used in the algorithm as its interpolation length satisfies the ODE $\dot{x} = f(x)$. However, to derive this, one has to *a priori* assume that $\{x_k\}$ is bounded. This assumption on $\{x_k\}$ in essence is a growth rate restriction on $f(\cdot)$.

SA methods are now widely applied in system identification, adaptive control, optimization, signal processing and other areas (See, e.g, [6,7] among others). However, the above-mentioned restrictions limit its further applications. To improve its applicability, the expanding truncation

[1] This work is supported by the National Key Project of China and by the National Natural Science Foundation of China.

technique was proposed in [8], and further developed in [9], [10], where the algorithm is analyzed by the trajectory-subsequence (TS) method, i.e., the analysis is carried out at a fixed trajectory along convergent subsequences.

In this paper the convergence theorems for SA algorithms with expanding truncations are presented. Then the method is applied to solve problems arising from optimization and signal processing.

2. SA ALGORITHM WITH EXPANDING TRUNCATIONS

Let $f(\cdot)$ be an $\mathbb{R}^l \to \mathbb{R}^l$ function with root set $J = \{x \in \mathbb{R} : f(x) = 0\}$. Let $\{M_k\}$ be a sequence of positive numbers increasingly diverging to infinity, and let x^0 be a fixed point in $\mathbb{R}^l$. Fix an arbitrary initial value x_0, and denote by x_k the estimate at time k serving as the k^{th} approach to J. Define x_k by the following algorithm with expanding truncations:

$$x_{k+1} = (x_k + a_k y_{k+1}) I_{[\|x_k + a_k y_{k+1}\| \le M_{\sigma_k}]} + x^0 I_{[\|x_k + a_k y_{k+1}\| > M_{\sigma_k}]} \quad (3)$$

$$\sigma_k = \sum_{i=1}^{k-1} I_{[\|x_i + a_i y_{i+1}\| > M_{\sigma_i}]}, \quad \sigma_0 = 0, \quad (4)$$

$$y_{k+1} = f(x_k) + \epsilon_{k+1}, \quad (5)$$

where $I_{[inequality]}$ is an indicator function meaning that it equals 1 if the inequality indicated in the bracket is fulfilled, and 0 if the inequality does not hold.

We explain the algorithm. σ_k is the number of truncations up-to time k. M_{σ_k} serves as the truncation bound when the $(k+1)th$ estimate is generated. From (3) it is seen that if the estimate at time $k+1$ calculated by the RM algorithm remains in the truncation region, i.e., if $\|x_k + a_k y_{k+1}\| \le M_{\sigma_k}$, then the algorithm evolves as the RM algorithm. If $(x_k + a_k y_{k+1})$ exits from the sphere with radius M_{σ_k}, i.e., if $\|x_k + a_k y_{k+1}\| > M_{\sigma_k}$, then the estimate at time $k+1$ is pulled back to the pre-specified point x^0, and the truncation bound is enlarged from M_{σ_k} to $M_{\sigma_{k+1}}$. Consequently , if it can be shown that the number of truncations is finite, or equivalently, $\{x_k\}$ generated by (3)–(5) is bounded, then the algorithm (3)–(5) turns to the RM algorithm in a finite number of steps.

Before establishing convergence of x_k, it is even unknown if x_k is bounded or not. So, in the case where ϵ_{k+1} depends on $x_j, j \le k$ (state-dependent noise), it is difficult to analyze the properties of $\{\epsilon_{k+1}\}$ along the whole sequence $\{x_k\}$. The noise conditions required here are needed to be verified only along convergent subsequences of $\{x_k\}$, and the analysis is carried out at a fixed trajectory (sample path). This is why we call it as TS method.

We first list conditions to be used.

A1 $a_k > 0, \quad a_k \xrightarrow[k\to\infty]{} 0 \quad$ and $\quad \sum_{k=1}^{\infty} a_k = \infty$

A2 There is a continuously differentiable function (not necessarily being non-negative) $v(\cdot) : \mathbb{R}^l \to R$ such that

$$\sup_{\delta \le d(x,J) \le \Delta} f^T(x) v_x(x) < 0 \quad (6)$$

for any $\Delta > \delta > 0$, and $v(J) \triangleq \{v(x) : x \in J\}$ is nowhere dense, where J is the zero set of $f(\cdot)$, i.e., $f(x) = 0, \forall x \in J, d(x, J) = \inf_y\{\|x - y\| : y \in J\}$ and $v_x(\cdot)$ denotes the gradient of $v(\cdot)$. Further, x^0 used in (3) is such that $v(x) < \inf_{\|x\|=c_0} v(x)$ for some $c_0 > 0$ and $\|x^0\| < c_0$.

For introducing condition on noise let us denote by $(\Omega, \mathcal{F}, P)$ the probability space. Let $\epsilon_{k+1}(\cdot,\cdot) = (\mathbb{R}^l \times \Omega, \mathcal{B}^l \times \mathcal{F}) \to (\mathbb{R}^l \times \mathcal{B}^l)$ be a measurable function defined on the product space. Fixing a $\omega \in \Omega$ means that a sample path is under consideration. Let the noise ϵ_{k+1} in (5) be given by

$$\epsilon_{k+1} = \epsilon_{k+1}(x_k, \omega), \quad \omega \in \Omega.$$

Thus, the state-dependent noise is considered, and for a fixed x, $\epsilon_{k+1}(x, \omega)$ may be random.

A3 For the sample path ω under consideration for any sufficiently large integer $N(\ge N_0)$

$$\lim_{t\to 0} \limsup_{k\to\infty} \frac{1}{t} \Big\| \sum_{i=n_k}^{m(n_k,t_k)} a_i \epsilon_{i+1}(x_i(\omega), \omega) \cdot I_{[\|x_i(\omega)\| \le N]} \Big\| = 0, \quad \forall t_k \in [0, t] \quad (7)$$

for any $\{n_k\}$ such that $x_{n_k}(\omega)$ converges, where $x_i(\omega)$ denotes x_i given by (3)–(5) and valued at the sample path ω and

$$m(k,t) = max\{m : \sum_{i=k}^{m} a_i \le t\}. \quad (8)$$

A4 $f(\cdot)$ is measurable and locally bounded.

In the sequel, the algorithm (3)–(5) is considered for a fixed ω for which A3 holds, and ω in $x_i(\omega)$ is often suppressed if no confusion will be caused. It is worth noting that $a_{n_k}\epsilon_{n_k+1} \xrightarrow[k\to\infty]{} 0$ if $\{x_{n_k}\}$ converges. To see this it suffices to take $t_k = a_{n_k}$ in (7).

Theorem 1. Let $\{x_k\}$ be given by (3)–(5) for a given initial value x_0. Assume A1–A4 hold. Then, $d(x_k, J^*) \xrightarrow[k\to\infty]{} 0$ for the sample path ω for which (7) holds, where J^* is a connected subset contained in $\bar{J}$, the closure of J [9, 2]. ◇

If for the conventional (untruncated) RM algorithm

$$x_{k+1} = x_k + a_k y_{k+1}, \quad y_{k+1} = f(x_k) + \epsilon_{k+1} \quad (9)$$

it is *a priori* known that $\{x_k\}$ is bounded, then we have the following theorem.

Theorem 2. Assume A1–A4 hold, where in A2 "Further, x^0 used in (3) is such that $v(x) < \inf_{\|x\|=c_0} v(x)$ for some $c_0 > 0$ and $\|x^0\| < c_0$" is deleted. If $\{x_k\}$ produced by (9) is bounded, then $d(x_k, J^*) \xrightarrow[k\to\infty]{} 0$ for the sample path ω for which A3 holds, where J^* is a connected subset of $\bar{J}$. ◇

We now give convergence theorems under conditions with no $\{x_k\}$ involved. For this we first reformulate Theorem 1.

In lieu of A3 we introduce the following condition.

A5 For any sufficiently large integer $N(\geq N_0)$ there is a $\omega-$set Ω_N with $P\Omega_N = 1$ such that for any $\omega \in \Omega_N$

$$\lim_{t\to 0} \limsup_{k\to\infty} \frac{1}{t} \Big\| \sum_{i=n_k}^{m(n_k,t_k)} a_i \epsilon_{i+1}(x_i(\omega), \omega) \cdot I_{[\|x_i(\omega)\| \leq N]} \Big\| = 0, \quad \forall t_k \in [0, t) \quad (10)$$

for any $\{n_k\}$ such that $\{x_{n_k}\}$ converges.

Theorem 3. Assume A1, A2, A4 and A5 hold. Then $d(x_k, J^*) \xrightarrow[k\to\infty]{} 0$ *a.s.* for $\{x_k\}$ generated by (3)–(5) with a given initial value x_0, where J^* is a connected subset contained in $\bar{J}$, the closure of J. ◇

We now introduce a state-independent condition on noise.

A6 $(\epsilon_k(x, \omega), \mathcal{F}_k)$ is a martingale difference sequence for any $x \in \mathbb{R}^l$, and for some $p \in (1, 2]$

$$E(\|\epsilon_{k+1}(x, \omega)\|^p | \mathcal{F}_k) \triangleq \sigma_{k+1}(x) < \infty, \quad a.s. \quad \forall x, \quad (11)$$

$$\sup_k \sup_{\|x\| \leq N} \sigma_{k+1}(x) \triangleq \sigma(N) < \infty, \quad \forall N, \quad (12)$$

where $\{\mathcal{F}_k\}$ is a family of nondecreasing σ-algebras independent of x.

Theorem 4. Let $\{x_k\}$ be given by (3)–(5) for a given initial value. Assume A1, A2, A4 and A6 hold and $\sum_{k=1}^{\infty} a_k^p < \infty$ for p given in A6. Then $d(x_k, J^*) \xrightarrow[k\to\infty]{} 0$ *a.s.*, where J^* is a connected subset contained in $\bar{J}$. ◇

In applications it may happen that $f(\cdot)$ is not directly observed. Instead, the time-varying functions $g_k(\cdot)$ are observed, and the observations y_{k+1} may be made not at x_k, but at $x_k + r_k$, i.e., at x_k with bias r_k,

$$y_{k+1} = g_k(x_k + r_k) + \epsilon_{k+1}(x_k + r_k, \omega). \quad (13)$$

Theorem 5. Let $\{x_k\}$ be given by (3)–(5) for a given initial value. Assume that A1, A2, A4 and A6 hold and $\sum_{k=1}^{\infty} a_k^p < \infty$ for p given in A6. Further, assume $(r_k, \mathcal{F}_k)$ is an adapted sequence, $\{r_k\}$ is bounded by a constant and for any sufficiently large integer $N(\geq N_0)$ there exists Ω_N with $P\Omega_N = 1$ such that for any $\omega \in \Omega_N$

$$\lim_{t\to 0} \limsup_{k\to\infty} \frac{1}{t} \sum_{i=n_k}^{m(n_k,t_k)} a_i (g_i(x_i + r_i) - f(x_i)) \cdot I_{[\|x_i\| \leq N]} \| = 0, \quad \forall t_k \in [0, t]. \quad (14)$$

for any $\{n_k\}$ such that $\{x_{n_k}\}$ converges. Then, $d(x_k, J^*) \xrightarrow[k\to\infty]{} 0$ a.s., where J^* is a connected subset contained in $\bar{J}$. ◇

3. OPTIMIZATION BY SA

Let $L(\cdot)$ be an unknown $\mathbb{R}^l \to R$ function. Based on the noisy measurements of $L(\cdot)$ the well-known Kiefer-Wolfowitz(KW) algorithm [12] is used to find the local maximizers of $L(\cdot)$. In order to reduce the number of observations at each step the algorithms with randomized differences are introduced [13, 14]. For the conventional KW algorithm $2l$ measurements at each step are needed, while for the algorithms with randomized differences only 2 measurements per step are used.

In contrast to [15], here the measurement noise at time $k+1$ is allowed to depend on the points ($x_k + c_k\Delta_k$ and $x_k - c_k\Delta_k$ to be explained below) where the function $L(\cdot)$ is observed.

At each iteration, two measurements are taken:

$$y_{k+1}^+ = L(x_k + c_k\Delta_k) + \xi_{k+1}^+(x_k + c_k\Delta_k, \omega) \quad (15)$$

$$y_{k+1}^- = L(x_k - c_k\Delta_k) + \xi_{k+1}^-(x_k - c_k\Delta_k, \omega) \quad (16)$$

where $\Delta_k \triangleq [\Delta_k^1, \ldots, \Delta_k^l]^T$ is defined as follows:

$(\Delta_k^i, i = 1, \ldots, l,\ k = 1, 2, \ldots)$ is a sequence of mutually independent and identically distributed random variables with $|\Delta_k^i| < a$, $|1/\Delta_k^i| < b$ and $E(1/\Delta_k^i) = 0$ for all $i \in \{1, \ldots, l\}$, $k = 1, 2, \ldots$, where a and b are positive real numbers.

$(\Delta_k^i, i = 1, \ldots, l, k = 1, 2, \ldots,)$ is independent of both sequences $\{\xi_k^+(x, \omega)\}$ and $\{\xi_k^-(x, \omega)\}$, $\forall x \in \mathbb{R}^l$.

It is important to note that both ξ_{k+1}^- and ξ_{k+1}^+ depend on Δ_k.

Define

$$y_{k+1} = \frac{(y_{k+1}^+ - y_{k+1}^-)}{2c_k} \Delta_k^{-1}, \quad (17)$$

$$\xi_{k+1} = \xi_{k+1}^+(x_k + c_k\Delta_k, \omega) - \xi_{k+1}^-(x_k - c_k\Delta_k, \omega), \quad (18)$$

where $\Delta_k^{-1} \triangleq [\frac{1}{\Delta_k^1}, \ldots, \frac{1}{\Delta_k^l}]^T$.

Let $\{x_k\}$ be still defined by (3)–(5) but with y_{k+1} given by (17).

We need the following conditions:

H1 The function $\nabla L = f$ is locally Lipschitz continuous. There is a unique maximum of L at x^0 so that $f(x^0) = 0$ and $f(x) \neq 0, \forall x \neq x^0$. There is a $c_0 > 0$ such that $\|x^0\| < c_0$ and $\sup_{\|x\|=c_0} L(x) < L(x^0)$ where x^0 is the one used in (3).

H2 $a_k > 0$, $c_k > 0$, $c_k \to 0$ as $k \to \infty$, $\sum_{k=1}^{\infty} a_k = \infty$ and there is a $p \in (1, 2]$ such that $\sum_{k=1}^{\infty} (\frac{a_k}{c_k})^p < \infty$.

H3 Both $\xi_k^+(x,\omega)$ and $\xi_k^-(x,\omega)$ are measurable functions: $(\mathbb{R}^l \times \Omega, \mathcal{B}^l \times \mathcal{F}) \to (R, \mathcal{B})$, $\forall k$, satisfying the following conditions

$$\xi_k^+(x,\omega) \in \mathcal{F}_k, \quad \xi_k^-(x,\omega) \in \mathcal{F}_k,$$

$$E(\xi_{k+1}^+(x,\omega)|\mathcal{F}_k) = 0, \quad E(\xi_{k+1}^-(x,\omega)|\mathcal{F}_k) = 0,$$

$$\forall x \in \mathbb{R}^l$$

$$E(\|\xi_{k+1}^+(x,\omega)\|^p|\mathcal{F}_k) \triangleq \sigma_{k+1}^+(x) < \infty,$$

$$E(\|\xi_{k+1}^-(x,\omega)\|^p|\mathcal{F}_k) \triangleq \sigma_{k+1}^-(x) < \infty,$$

$$\sup_k \sup_{\|x\| \le N} [\sigma_{k+1}^+(x) + \sigma_{k+1}^-(x)] \triangleq \sigma(N) < \infty, \quad \forall N, \tag{19}$$

where $\{\mathcal{F}_k\}$ is a family of nondecreasing σ-algebras independent of both x and $\{\Delta_k^i, i = 1, \ldots, l, k = 1, 2, \ldots\}$.

Theorem 6. Assume H1, H2 and H3 hold. Then $\{x_k\}$ defined by (3)–(5) with y_{k+1} given by (17) converges to x^0 a.s. ◇

4. ADAPTIVE FILTERING

Let $\{X_n\} \in \mathbb{R}^r$ and $\{s_n\} \in R$ be the measured input and reference signal, respectively. Assume $[s_n, X_n^T]^T$ is stationary. The vector H is adaptively adjusted so that $H^T X_n$ best matches s_n in a certain sense. If the cost function $L(H)$ to be minimized is $E|s_n - H^T X_n|$, then the following sign-algorithm

$$H_{k+1} = H_k + a_k X_k \text{sign}(s_k - H_k^T X_k), \quad a_k = \frac{1}{k} \tag{20}$$

is used to estimate the minimizer of $L(H)$. If the cost function $L(H) = E|s_n - H^T X_n|^2$, then the following algorithm

$$H_{k+1} = H_k + a_k X_k (s_k - H_k^T X_k) \tag{21}$$

has extensively been studied.

According to (3)–(5), instead of (20),(21) we consider the corresponding versions with expanding truncations.

Take $\{M_k\}$, $M_k > M_{k-1}, \forall k$, $M_k \to \infty$ as $k \to \infty$. The algorithms (20), (21) are modified to (22), (23), respectively

$$H_{k+1} = (H_k + a_k X_k \text{sign}(s_k - H_k^T X_k)) \cdot I_{\{\|H_k + a_k X_k \text{sign}(s_k - H_k^T X_k)\| \le M_{\sigma_k}\}} \tag{22}$$

$$\sigma_k = \sum_{i=1}^{k-1} I_{\{\|H_i + a_i X_i \text{sign}(s_i - H_i^T X_i)\| > M_{\sigma_i}\}}, \quad a_k = \frac{1}{k}$$

and

$$H_{k+1} = (H_k + a_k X_k (s_k - H_k^T X_k)) \cdot I_{\{\|H_k + a_k X_k (s_k - H_k^T X_k)\| \le M_{\sigma_k}\}} \tag{23}$$

$$\sigma_k = \sum_{i=1}^{k-1} I_{\{\|H_i + a_i X_i (s_i - H_i^T X_i)\| > M_{\sigma_i}\}}, \quad a_k = \frac{1}{k}.$$

Notice that x^0 in (3) is now set to be zero in (22) and (23).

Theorem 7. Assume $[s_k, X_k^T]^T$ is stationary and ergodic such that

$$E \begin{bmatrix} s_1 \\ X_1 \end{bmatrix} [s_1 X_1^T] > 0.$$

Then as $k \to \infty$, $d(H_k, J) \to 0$ a.s. for H_k defined by both (22) and (23), where J denotes the set of minimizers of $E|s_1 - H^T X_1|$ for (22) and $E|s_1 - H^T X_1|^2$ for (23). ◇

It is clear that to minimize $L(H) = E(s_1 - H^T X_1)^2$ is equivalent to finding the roots of its gradient

$$f(H) = E X_1^T (s_1 - H^T X_1).$$

Therefore, the corresponding RM algorithm should be

$$H_{k+1} = H_k - a_k E(X_1(s_1 - H_k^T X_1)) \quad (= H_k - a_k f(H_k)). \tag{24}$$

Comparing (24) with (21) we rewrite (21) as

$$H_{k+1} = H_k - a_k f(H_k) + a_k \epsilon_{k+1}, \tag{25}$$

where

$$\epsilon_{k+1} = X_k(s_k - H_k^T X_k) + E(X_1(s_1 - H_k^T X_1)). \tag{26}$$

Thus, (25) is a RM algorithm with observation noise (26), which is state-dependent. The similar situation takes place when minimizing $E|s_n - H^T X_n|$, for which

$$f(H) = E(X_1 sign(s_1 - H^T X_1))$$

and

$$\epsilon_{k+1} = X_k sign(s_k - H_k^T X_k) + E(X_1 sign(s_1 - H_k^T X_1)).$$

The proof of Theorem 7 consists in verifying condition A3, applying ergodicity of the stationary process $(s_k, X_k^T)^T$. Then the assertions of the theorem follows from Theorem 1. For details we refer to [16]. Comparing Theorem 7 with results given in [17] we find that conditions used in [17] have greatly been weakened.

5. BLIND CHANNEL IDENTIFICATION

Consider a system consisting of p FIR channels with L being the maximum order of the channels. Let s_k, $k = 0, 1, 2, \cdots, N$, be the one-dimensional input signal, and $x_k = (x_k^{(1)}, x_k^{(2)}, \cdots, x_k^{(p)})^T$, $k = L, L+1, \cdots, N$, be the p-dimensional output signal, where N is the number of samples and may not be fixed, the superscript (i) denotes the ith component, and the subscript k is the time index. Then

$$x_k = \sum_{i=0}^{L} h_i s_{k-i}, \qquad k \geq L, \tag{27}$$

where

$$h_i = [h_i^{(1)}, \cdots, h_i^{(p)}]^T.$$

Equation (27) can be written as

$$x_k^{(i)} = h^{(i)}(z) s_k, \tag{28}$$

where

$$h^{(i)}(z) \triangleq h_0^{(i)} + h_1^{(i)} z + \cdots + h_L^{(i)} z^L, \quad i = 1, \ldots, p, \tag{29}$$

with z being the shift operator:

$$z s_k = s_{k-1}.$$

The observations y_k may be corrupted by noise n_k:

$$y_k = x_k + n_k,$$

where n_k is a p-dimensional noise vector. The problem is to estimate h_i, $i = 0, \cdots, L$, on the basis of observations $\{y_k\}$. Note that s_k, x_k, n_k, and y_k can be complex numbers.

The channels can be characterized by a $p(L+1)$-dimensional vector h^0. First we define

$$h^{(i)} = (h_0^{(i)}, \cdots, h_L^{(i)})^T,$$

then let

$$h^0 = \left[(h^{(1)})^T, \cdots, (h^{(p)})^T\right]^T. \tag{30}$$

Denote

$$\psi_k^{(i)} = \left[y_k^{(i)} \cdots y_{k-L}^{(i)}\right], \ \varphi_k^{(i)} = \left[x_k^{(i)} \cdots x_{k-L}^{(i)}\right], \quad i = 1, \cdots, p, \ k \geq 2L,$$

where $y_k^{(i)}$ and $x_k^{(i)}$ are the ith component of y_k and x_k, respectively.

¿From (28), we have

$$\begin{aligned} & h^{(i)}(z) x_k^{(j)} \\ =& h^{(i)}(z) h^{(j)}(z) s_k = h^{(j)}(z) h^{(i)}(z) s_k = h^{(j)}(z) x_k^{(i)}, \\ & \forall i, j = 1, \ldots, p, k = 2L, 2L+1, \ldots. \end{aligned} \tag{31}$$

Using the observed data (y_k or x_k in the noise free case), the above set of equations can be written in a matrix form [18]

$$X_L h^0 = 0, \tag{32}$$

where X_L is a $(N - 2L + 1)[p(p-1)/2] \times [(L+1)p]$ matrix, and N is the number of samples. In all existing results the "block algorithm" is used, where the channel coefficients are estimated after the entire block of data have been received. In contrast to this, by using the SA method we propose adaptive algorithms in which estimates for h^0 are obtained at every step $k = 2L, 2L+1, \cdots, N$, by updating the estimates obtained at the previous step.

First, we define two $\frac{p(p-1)}{2} \times p(L+1)$ matrices denoted as Ψ_k and Φ_k:

$$\Psi_k = \begin{bmatrix} \psi_k^{(2)} & -\psi_k^{(1)} & 0 & \cdots & \cdots & 0 \\ \psi_k^{(3)} & 0 & -\psi_k^{(1)} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \psi_k^{(p)} & \psi_k^{(2)} & 0 & \cdots & 0 & -\psi_k^{(1)} \\ 0 & \psi_k^{(3)} & -\psi_k^{(2)} & 0 & \cdots & 0 \\ & \vdots & \vdots & \ddots & \ddots & \\ 0 & \psi_k^{(p)} & 0 & \cdots & 0 & -\psi_k^{(2)} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & \vdots & & & \vdots & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & 0 & \psi_k^{(p)} & -\psi_k^{(p-1)} \end{bmatrix}, \tag{33}$$

Note that X_L in (32) is $(N - 2L + 1)$ times as large as $\Psi_k^{(i)}$. We define Φ_k as the matrix that has the same structure as (33) with $\psi_k^{(i)}$ replaced by $\varphi_k^{(i)}$ $\forall i = 1, 2, \cdots, p$. Ψ_k (or Φ_k) contains the observation x_k in the noise-free case (or y_k for noisy observations) in a window of size $L+1$ back from time instant k (i.e., $k, k-1, \cdots, k-L$); these are the observations that are related to signal s_{k-L}. It is worth emphasizing that neither Ψ_k and Φ_k depend on N in contrast to X_L.

Let $\{a_k\}$ be a sequence of step sizes to be specified later. Let $\|h(2L-1)\| < \kappa$ and estimate h^0 by the following truncated SA algorithm:

$$\begin{aligned} & h(k+1) \\ =& \left(h(k) - a_k \left(\Psi_{k+1}^* \Psi_{k+1} - E N_{k+1}^* N_{k+1}\right) h(k)\right) \\ & \cdot I_{[\|h(k) - a_k(\Psi_{k+1}^* \Psi_{k+1} - E N_{k+1}^* N_{k+1})\| < 1]} \\ & + h(2L-1) I_{[\|h(k) - a_k(\Psi_{k+1}^* \Psi_{k+1} - E N_{k+1}^* N_{k+1})\| \geq 1]}, \\ & k = 2L, 2L+1, \cdots, \end{aligned} \tag{34}$$

where the superscript $*$ denotes transpose with complex conjugate, and $N_k = \Psi_k - \Phi_k$, $k = 2L, 2L+1, \cdots$.

We will use the following conditions.

C1 $h^i(z)$, $i = 1, \cdots, p$, given by (29) have no common factor.

C2 $a_k > 0$, $a_{k+1} \leq a_k$ $\forall k = 1, 2, \cdots$, $a_k \underset{k \to \infty}{\to} 0$, $\sum_{k=1}^{\infty} a_k = \infty$, $\frac{a_k}{a_{k+1}} < c$ $\forall k$ with c

being a constant and $\sum_{i=1}^{\infty} a_i^{1+\frac{\gamma}{2}} < \infty$, where γ is given in C3.

C3 $\{s_k\}$ and $\{n_k\}$ are mutually independent and each of them is a sequence of mutually independent random variables such that $E|s_k|^2 \neq 0$, and

$$\sup_k \{|s_k| + |n_k|\} \leq \eta < \infty, \quad E\eta^{2+\gamma} < \infty.$$

C4 $\lambda_{\min}(j,k) \geq \lambda > 0$, $\forall j \geq 0$, $\forall k \geq 0$, where $\lambda_{\min}(j,k)$ is the minimal nonzero eigenvalue of $B_{j,k}$, where

$$B_{j,k} \triangleq \sum_{i=j+k(2L+1)}^{j+(k+1)(2L+1)-1} E\Phi_i^*\Phi_i, \quad \forall j \geq 0, \quad \forall k \geq 0. \tag{35}$$

Theorem 8. Assume C1–C4 hold, and $h(k)$ is given by (34) with initial value $h(2L-1)$. Then after a finite number of steps there is no truncation in (34) and

$$h(k) \underset{k\to\infty}{\to} \alpha h^0, \qquad \text{a.s.}$$

where α is a random variable. ◇

Note that $\bar{h^0} = h^0/\|h^0\|$ is the unique unit eigenvector for $B_{j,k}, \forall j \geq 0, \forall k \geq 0$, and (34) is a SA algorithm but with time-varying regression function. For the proof we refer to [19].

6. REFERENCES

[1] Robbins, H. and Monro, S., A stochastic approximation method, *Ann. Math, Statist.*, Vol. 22, 1951, 400–407.

[2] Fang, H.T. and Chen, H.F., Stability and Instability of limit points for stochastic approximation algorithms, publication in *IEEE Trans. Autom. Control,*Vol. 45, No. 3, 2000, 413–420.

[3] Nevelson, M.B. and Hasminskii, R.A., *Stochastic Approximation and Recursive Estimation,* AMS Transl. Math. Monographs, 49, 1976.

[4] Ljung, L., Analysis of recursive stochastic algorithms, *IEEE Trans. Autom. Control,* Vol. 22, 1977, 551–575.

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

[6] Benveniste, A., Métivier, M. and Priouret, P., *Adaptive Algorithms and Stochastic Approximations,* Springer-Verlag, 1990, Berlin.

[7] Kushner, H.J. and Yin, G., *Stochastic Approximation Algorithms and Applications,* Springer-Verlag, Berlin, 1997.

[8] Chen, H.F. and Zhu, Y.M., Stochastic approximation procedures with randomly varying truncations, Scientia Sinica (Series A), Vol. 29, No. 9, 1986, 914-926.

[9] Chen, H.F., Stochastic approximation and its new applications, in *Proc. Hong Kong Int. Workshop on New Directions of Control and Manufacturing,* 1994, 2–12.

[10] Chen, H. F., *Stochastic Approximation and Its Applications,* Kluwer, to appear in 2002.

[11] Chen, H. F., Stochastic Approximation with State-Dependent Noise, *Science in China (Series E),* Vol. 43, No. 5, 2000, 531–541.

[12] Kiefer, J. and Wolfowitz, J., Stochastic approximation of a regression function, *Ann. Math. Stat.,* Vol. 23, 1952, 552–558.

[13] Koronaski, J., Random-seeking methods for the stochastic unconstrained optimization, *Int. J. Control,* Vol. 21, 1975, 517–527.

[14] Spall, J.C., Multivariate stochastic approximation using a simultaneous perturbation gradient approximation, *IEEE Trans. Autom. Control,* Vol. 37, 1992, 332–341.

[15] Chen, H.F., Duncan, T.E. and Pasik-Duncan, B., A Kiefer-Wolfowitz algorithm with randomized differences, *IEEE Trans. Autom. Control,* Vol. 44, No. 3, 1999, 442–453.

[16] Chen, H. F. and G. Yin, Asymptotic properties of sign algorithms for adaptive filtering, submitted for publication.

[17] Eweda, E., . Convergence of the sign algorithm for adaptive filtering with correlated data, *IEEE Trans. Inform. Theory,* IT-37, 1991, 1450–1457.

[18] Xu, G., L. Tong and T. Kailath, A least squares approach to blind identification, *IEEE Trans. Signal Processing,* Vol. 43, No. 12, 1995, 2982-2993.

[19] Chen, H. F., X. R. Cao and J. Zhu, Convergence of stochastic approximation based algorithms for blind channel identification, submitted for publication.