STOCHASTIC APPROXIMATION ALGORITHMS
WITH EXPANDING TRUNCATIONS

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 \( f(\cdot) \):

\[ x_{k+1} = x_k + a_k y_{k+1} \]

(1)

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

\[ y_{k+1} = f(x_k) + \epsilon_{k+1} \]

(2)

where \( \epsilon_{k+1} \) 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
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 \( \mathbb{R}^{d} \rightarrow \mathbb{R}^{d} \) function with root set \( J = \{ x \in \mathbb{R} : f(x) = 0 \} \). Let \( \{ M_{k} \} \) be a sequence of positive numbers increasing to infinity, and let \( x^{0} \) be a fixed point in \( \mathbb{R}^{d} \). 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:

\[
\begin{align*}
\sigma_{k+1} &= \sigma_{k} + a_{k}y_{k+1}, \\
\sigma_{0} &= 0, \\
y_{k+1} &= f(x_{k}) + \varepsilon_{k+1}, \\
\end{align*}
\]

where \( I_{[\text{inequality}]} \) 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. \( \sigma_{k} \) is the number of truncations up-to time \( k \). \( M_{k} \) 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 \( \|x_{k} + a_{k}y_{k+1}\| \leq M_{k} \), then the algorithm evolves as the RM algorithm. If \( \|x_{k} + a_{k}y_{k+1}\| \) exceeds the sphere with radius \( M_{k} \), i.e., if \( \|x_{k} + a_{k}y_{k+1}\| > M_{k} \), then the estimate at time \( k+1 \) is pulled back to the pre-specified point \( x^{0} \), and the truncation bound is enlarged from \( M_{k} \) to \( M_{k+1} \). 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 \( \varepsilon_{k+1} \) depends on \( x_{i} \), \( i \leq k \) (state-dependent noise), it is difficult to analyze the properties of \( \{\varepsilon_{k+1}\} \) 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** \( a_{k} > 0, \quad a_{k} \xrightarrow[k \to \infty]{} 0 \) and \( \sum_{k=1}^{\infty} a_{k} = \infty \)

**A2** There is a continuously differentiable function (not necessarily being non-negative) \( \psi(\cdot) : \mathbb{R}^{d} \rightarrow \mathbb{R} \) such that
\[
\sup_{\delta \leq \|x_{i},J\| \leq \delta} f^{T}(x_{i})u_{x}(x) < 0
\]
for any \( \Delta > \delta > 0 \), and \( \psi(J) \uparrow \{\psi(x) : x \in J\} \) is nowhere dense, where \( J \) is the zero set of \( f(\cdot) \), i.e., \( f(x) = 0, \forall x \in J, d(x, J) = \inf_{y \in J} \|x - y\| : y \in J \) and \( u_{x}(\cdot) \) denotes the gradient of \( \psi(\cdot) \). Further, \( \delta^{0} \) used in (3) is such that \( \psi(x) < \inf_{\|x\| = \delta_{0}} \psi(x) \) for some \( \delta_{0} > 0 \) and \( \|x^{0}\| < \delta_{0} \).

For introducing condition on noise we list denote by \( (\Omega, F, F) \) the probability space. Let \( \varepsilon_{k+1}(\cdot, \cdot) = (\mathbb{R}^{d} \times \Omega, \mathcal{B}^{d} \times F) \rightarrow (\mathbb{R}^{d} \times \mathcal{B}^{d}) \) be a measurable function defined on the product space. Fixing \( \omega \in \Omega \) means that a sample path is under consideration. Let the noise \( \varepsilon_{k+1} \) in (5) be given by
\[
\varepsilon_{k+1} = \varepsilon_{k+1}(x_{k}, \omega), \quad \omega \in \Omega.
\]

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

**A3** For the sample path \( \omega \) under consideration for any sufficiently large integer \( N(\geq N_{0}) \)
\[
\lim_{t \to 0} \sup_{k \to \infty} \frac{1}{t} \sum_{i=0}^{m(k,t)} a_{i} \varepsilon_{i+1}(x_{i}(\omega), \omega)
\]
\[
-I_{[\|x_{i}(\omega)\| \leq \|x\|]}(x_{i}(\omega), \omega) = 0, \quad \forall t \in [0, t]
\]
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 \( \omega \) and
\[
m(k, t) = \max\{m : \sum_{i=0}^{m} a_{i} \leq t\}.
\]

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

In the sequel, the algorithm (3)–(5) is considered for a fixed \( \omega \) for which A3 holds, and \( \omega \) in \( x_{i}(\omega) \) is often suppressed if no confusion will be caused. It is worth noting that \( a_{n_{k}} \varepsilon_{n_{k}+1} \xrightarrow[k \to \infty]{} 0 \) if \( \{x_{k}\} \) 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, \( d(x_{k}, J^{\ast}) \xrightarrow[k \to \infty]{} 0 \) for the sample path \( \omega \) for which (7) holds, where \( J^{\ast} \) is a connected subset contained in \( J \), the closure of \( J \) [9, 2].

If for the conventional (untruncated) RM algorithm
\[ z_{k+1} = z_k + a_k y_{k+1}, \quad y_{k+1} = f(z_k) + \epsilon_{k+1} \quad (9) \]
it is a priori known that \( \{z_k\} \) 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 \( \omega \) for which A3 holds, where \( J^* \) is a connected subset of \( J \).

We now give convergence theorems under conditions with no \( \{z_k\} \) 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 \( \Omega_N \) with \( P \Omega_N = 1 \) such that for any \( \omega \in \Omega_N \)

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

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 \( J \), the closure of \( J \).

We now introduce a state-independent condition on noise.

**A6** \((\epsilon_k(x, \omega), F_k)\) is a martingale difference sequence for any \( x \in \mathbb{R}^l \), and for some \( p \in (1, 2) \)

\[
E(||\epsilon_k(x, \omega)||^p | F_k) \triangleq \sigma(x_k < \infty, \quad a.s., \quad \forall x, \quad (11)
\]

\[
sup_{k} sup_{x \in X} \sigma(x_{k+1}) \geq \sigma(N) < \infty, \quad \forall N, \quad (12)
\]

where \( \{F_k\} \) is a family of nondecreasing \( \sigma \)-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 \( J \).

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 \( (r_k, F_k) \) is an adapted sequence, \( \{r_k\} \) is bounded by a constant and for any sufficiently large integer \( N \geq N_0 \) there exists \( \Omega_N \) with \( P \Omega_N = 1 \) such that for any \( \omega \in \Omega_N \)

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

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 \( J \).

3. OPTIMIZATION BY SA

Let \( L(\cdot) \) be an unknown \( \mathbb{R}^d \rightarrow \mathbb{R} \) 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:

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

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

It is important to note that both \( \xi_{k+1}^- \) and \( \xi_{k+1}^+ \) depend on \( \Delta_k \).

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 \( \Delta_k^{-1} \triangleq [\Delta_k^{-1}, \ldots, \Delta_k^{-1}]^T. \)
Let \( \{z_k\} \) 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 \( k \to \infty \), \( \sum_{k=1}^{\infty} c_k = \infty \) and there is a \( p \in (1,2) \) such that 
\[ \sum_{k=1}^{\infty} (\frac{c_k}{k})^p < \infty. \]

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

\[ E(\xi^+ (x, \omega) | F_k) = 0, \quad E(\xi^- (x, \omega) | F_k) = 0, \]
\[ \forall x \in \mathbb{R}^d \]
\[ E(||\xi^+ (x, \omega)||^p | F_k) \leq \sigma^+ (x) < \infty, \]
\[ E(||\xi^- (x, \omega)||^p | F_k) \leq \sigma^- (x) < \infty, \]
\[ \sup_k \sup_{||x|| \leq N} (\sigma^+ (x) + \sigma^- (x)) \leq \sigma (N) < \infty, \]
\[ \forall N. \] \hspace{1cm} (19)

where \( \{F_k\} \) is a family of nondecreasing \( \sigma \)-algebras independent of both \( x \) and \( \Delta_k, i = 1, \ldots, t, k = 1, 2, \ldots \).

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

\[ \Diamond \]

4. ADAPTIVE FILTERING

Let \( \{X_n\} \in \mathbb{R}^r \) and \( \{s_n\} \in \mathbb{R} \) 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

\[ H_{k+1} = H_k + a_k X_k \text{sign}(s_k - H_k^T X_k), \quad a_k = \frac{1}{k}. \] \hspace{1cm} (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

\[ H_{k+1} = H_k + a_k X_k (s_k - H_k^T X_k) \] \hspace{1cm} (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

\[ H_{k+1} = (H_k + a_k X_k \text{sign}(s_k - H_k^T X_k)) \]
\[ I(||H_k + a_k X_k \text{sign}(s_k - H_k^T X_k)|| \leq M_k) \]
\[ \sigma_k = \sum_{i=1}^{k-1} I(||H_i + a_i X_i \text{sign}(s_i - H_i^T X_i)|| \leq M_{\text{max}}), \quad a_k = \frac{1}{k} \]

and

\[ H_{k+1} = (H_k + a_k X_k (s_k - H_k^T X_k)) \]
\[ I(||H_k + a_k X_k (s_k - H_k^T X_k)|| \leq M_k) \]
\[ \sigma_k = \sum_{i=1}^{k-1} I(||H_i + a_i X_i (s_i - H_i^T X_i)|| \leq M_{\text{max}}), \quad a_k = \frac{1}{k}. \]

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

**Theorem 7.** Assume \( \{s_n, X_n^T\}^T \) is stationary and ergodic such that

\[ E\left[ \frac{s_1}{X_1} \right] \frac{s_1 X_1^T}{X_1} > 0. \]

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).

\[ \Diamond \]

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

\[ f(H) = EX_1^T(s_1 - H^T X_1). \]

Therefore, the corresponding RM algorithm should be

\[ H_{k+1} = H_k - a_k f(H_k) + a_k \epsilon_{k+1}, \]
\[ \epsilon_{k+1} = X_k (s_k - H_k^T X_k) + E(X_1 (s_1 - H_1^T X_1)). \] \hspace{1cm} (24)

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

\[ f(H) = E(X_1 \text{sign}(s_1 - H^T X_1)) \]

and

\[ \epsilon_{k+1} = X_k \text{sign}(s_k - H_k^T X_k) + E(X_1 \text{sign}(s_1 - H_1^T X_1)). \]

The proof of Theorem 7 consists in verifying condition A3, applying ergodicity of the stationary process \( \{s_n, X_n^T\}^T \). 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 \( s_k, k = 0, 1, 2, \ldots, N, \) be the one-dimensional input signal, and \( x_k = (x_k^{(1)}, x_k^{(2)}, \ldots, x_k^{(p)})^T, \quad k = L, L + 1, \ldots, N, \) be the \( p \)-dimensional output signal, where \( N \) is the number of samples and may not be fixed, the superscript \((i)\) denotes the \( i \)-th component, and the subscript \( k \) is the time index. Then

\[
x_k = \sum_{i=0}^{L} h_i s_{k-i}, \quad k \geq L, \tag{27}
\]

where

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

Equation (27) can be written as

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

where

\[
h^{(i)}(z) \overset{\Delta}{=} h_i^{(1)} + h_i^{(2)} z + \cdots + h_i^{(p)} z^L, \quad i = 1, \ldots, p,
\]

with \( z \) being the shift operator:

\[
s_k - s_{k-1}.
\]

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 \( h_i, i = 0, \ldots, L, \) 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

\[
h^{(i)} = (h_i^{(1)}, \ldots, h_i^{(p)})^T,
\]

then let

\[
h^0 = [(h^{(1)})^T, \ldots, (h^{(p)})^T]^T. \tag{30}
\]

Denote

\[
\psi_k^{(i)} = [y_k^{(i)} \ldots y_{k-L}^{(i)}], \quad \phi_k^{(i)} = [x_k^{(i)} \ldots x_{k-L}^{(i)}],
\]

where \( y_k^{(i)} \) and \( x_k^{(i)} \) are the \( i \)-th component of \( y_k \) and \( x_k \), respectively.

From (28), we have

\[
h^{(i)}(z)x_k^{(j)} \overset{\Delta}{=} h^{(i)}(z)s_k = h^{(j)}(z)x_k^{(i)} = h^{(j)}(z)x_k^{(i)},
\]

\[
\forall i, j = 1, \ldots, p, k = 2L, 2L+1, \ldots, \tag{31}
\]

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 \( k = 2L, 2L + 1, \ldots, N, \) 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 \( \Psi_k \) and \( \Phi_k: \)

\[
\Psi_k = \begin{bmatrix}
\psi_k^{(2)} - \psi_k^{(1)} & 0 & \ldots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & \ldots & -\psi_k^{(1)} & 0 \\
0 & \ldots & \ldots & 0 \\
0 & \ldots & \ldots & \psi_k^{(p)} \\
0 & \ldots & \ldots & \psi_k^{(p)} \\
\end{bmatrix},
\]

(33)

\[
\Phi_k = \begin{bmatrix}
\psi_k^{(2)} - \psi_k^{(1)} & 0 & \ldots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & \ldots & -\psi_k^{(1)} & 0 \\
0 & \ldots & \ldots & 0 \\
0 & \ldots & \ldots & \psi_k^{(p)} \\
0 & \ldots & \ldots & \psi_k^{(p)} \\
\end{bmatrix},
\]

\[
\Psi_k \overset{\Delta}{=} \begin{bmatrix}
\psi_k^{(2)} - \psi_k^{(1)} & 0 & \ldots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & \ldots & -\psi_k^{(1)} & 0 \\
0 & \ldots & \ldots & 0 \\
0 & \ldots & \ldots & \psi_k^{(p)} \\
0 & \ldots & \ldots & \psi_k^{(p)} \\
\end{bmatrix}.
\]

Note that \( X_L \) in (32) is \((N - 2L + 1)\) times as large as \( \Psi_k^{(i)} \). We define \( \Phi_k \) as the matrix that has the same structure as (33) with \( \psi_k^{(i)} \) replaced by \( \varphi_k^{(i)} \) \( \forall i = 1, 2, \ldots, p. \) \( \Psi_k \) (or \( \Phi_k \)) 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., \( k, k - 1, \ldots, k - L \)); these are the observations that are related to signal \( s_{k-L} \). It is worth emphasizing that neither \( \Psi_k \) and \( \Phi_k \) 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:

\[
h(k + 1) = (h(k) - a_k \left( \Psi_{k+1}^{-1} \Psi_{k+1}^{(1)} - EN_{k+1}^{(1)} N_{k+1} \right) h(k))
\]

\[
\cdot \left( I(h(k) - a_k \left( \Psi_{k+1}^{(1)} \Psi_{k+1}^{(1)} - EN_{k+1}^{(1)} N_{k+1} \right) ||<1 \right)
\]

\[
+ h(2L-1)I(h(k) - a_k \left( \Psi_{k+1}^{(1)} \Psi_{k+1}^{(1)} - EN_{k+1}^{(1)} N_{k+1} \right) || \geq 1), \tag{34}
\]

\[
k = 2L, 2L + 1, \ldots,
\]

where the superscript \( * \) denotes transpose with complex conjugate, and \( N_k = \Psi_k - \Phi_k, k = 2L, 2L + 1, \ldots, \)

We will use the following conditions.

\[\text{C1} \quad h^{(i)}(z), \quad i = 1, \ldots, p, \text{ given by (29) have no common factor.}\]

\[\text{C2} \quad a_k > 0, \quad a_{k+1} \leq a_k \quad \forall k = 1, 2, \ldots, \quad a_k \rightarrow 0, \quad \sum_{k=1}^{\infty} a_k = \infty, \quad \frac{a_k}{a_{k+1}} < c \quad \forall k \quad \text{with } c < 1.\]
being a constant and $\sum_{i=1}^{\infty} a_i^{1+\frac{\gamma}{2}} < \infty$, where $\gamma$ is given in C3.

C3 $\{x_k\}$ 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

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

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

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

(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

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

where $\alpha$ is a random variable.

Note that $h^0 = h^0/\|h^0\|$ 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