Convergence Behavior of Affine Projection Algorithms

Sundar G. Sankaran, *Student Member, IEEE*, and A. A. (Louis) Beex, *Senior Member, IEEE*

***Abstract*—Over the last decade, a class of equivalent algorithms that accelerate the convergence of the normalized LMS (NLMS) algorithm, especially for colored inputs, has been discovered independently. The affine projection algorithm (APA) is the earliest and most popular algorithm in this class that inherits its name. The usual APA algorithms update weight estimates on the basis of multiple, unit delayed, input signal vectors. We analyze the convergence behavior of the generalized APA class of algorithms (allowing for arbitrary delay between input vectors) using a simple model for the input signal vectors. Conditions for convergence of the APA class are derived. It is shown that the convergence rate is exponential and that it improves as the number of input signal vectors used for adaptation is increased. However, the rate of improvement in performance (time-to-steady-state) diminishes as the number of input signal vectors increases. For a given convergence rate, APA algorithms are shown to exhibit less misadjustment (steady-state error) than NLMS. Simulation results are provided to corroborate the analytical results.**

I. Introduction

ADAPTIVE filtering techniques are used in a wide range of applications, including adaptive equalization, adaptive noise cancellation, echo cancellation, and adaptive beamforming. The normalized least mean square (NLMS) algorithm [1] is a widely used adaptation algorithm due to its computational simplicity and ease of implementation. Furthermore, this algorithm is known to be robust against finite word length effects. One of the major drawbacks of the NLMS algorithm is its slow convergence for colored input signals. Over the last decade, a class of equivalent algorithms such as the affine projection algorithm (APA), the partial rank algorithm (PRA), the generalized optimal block algorithm (GOBA), and NLMS with orthogonal correction factors (NLMS-OCF) has been developed to ameliorate this problem [2], [3]. The distinguishing characteristic of these algorithms, which was developed independently from different perspectives, is that they update the weights on the basis of multiple, delayed input signal vectors, whereas the NLMS algorithm updates the weights on the basis of a single input vector. In the sequel, we will refer to the entire class of algorithms as affine projection algorithms, since APA (with unit delayed input vectors) is the earliest among these algorithms and since the name APA is more widely used in the existing literature than the other names. However, the convergence results that we derive here are applicable to the entire class of affine projection algorithms, allowing for arbitrary delay between input vectors.

The APA is a better alternative than NLMS in applications where the input signal is highly correlated [9], [10], [15]. Although a wide range of analysis has been done on the convergence behavior of the NLMS algorithm [4], [5], the convergence behavior of APA has not received as much attention to date. Some results are available on the steady-state behavior (characterized by misadjustment) of APA [11]–[13]. In this discussion, we analyze the convergence behavior of APA and derive the necessary and sufficient conditions for the convergence of the APA class of algorithms, as well as an expression for the mean-squared error. Furthermore, we study the improvement in performance with the number of vectors used for adaptation. The steady-state behavior is also analyzed. The analysis is done using a simple model for the input signal vector. In addition to the usual independence assumption [1], the angular orientation of the input vectors is assumed to be discrete. Although these assumptions are rarely satisfied by real-life data, they render the convergence analysis tractable. Furthermore, we show that simulation results match our analytical results when the data ("pretty much") satisfies the independence assumption. The limitations imposed by the assumptions used, as well as by the simplifications made in our analysis, are also discussed. Not unexpectedly, our analytical results deviate from the simulation results when the data grossly violates the assumptions; however, the general performance characteristics predicted by our analysis still hold. Thus, our results serve as useful design guidelines.

The weight update equation of APA is presented in Section II. Section III begins with a list of the assumptions that are used. Based on these assumptions, the convergence behavior of APA is analyzed. The insights provided by the analytical results are summarized. Section IV compares our analytical results with the results obtained from simulations. A summary of the results and concluding remarks are provided in Section V.

Notations used in this paper are fairly standard. Boldface symbols are used for vectors (in lowercase letters) and matrices (in uppercase letters). We also have the following notations:

- $(\cdot)^t$ transpose;
- $(\cdot)^H$ Hermitian transpose;
- $(\cdot)^*$ complex conjugate;
- $P(\cdot)$ probability;
- $E(\cdot)$ expectation;
- $\mathrm{tr}(\cdot)$ trace.

Manuscript received March 31, 1998; revised September 23, 1999. The associate editor coordinating the review of this paper and approving it for publication was Prof. James A. Bucklew.

The authors are with the Systems Group—DSP Research Laboratory, Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24061-0111 USA (e-mail: beex@vt.edu).

Publisher Item Identifier S 1053-587X(00)02358-8.

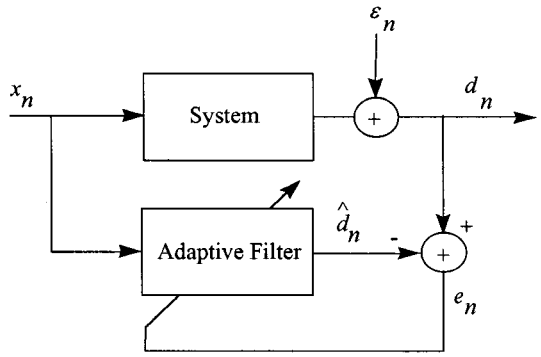


Fig. 1. Adaptive filtering problem.

II. Class of Affine Projection Algorithms

Fig. 1 shows an adaptive filter used in the system identification mode. Here, the system input x_n and corresponding measured output d_n, possibly contaminated with measurement noise ε_n, are known. The objective is to estimate an N-dimensional weight vector $\mathbf{w}_n$ such that the estimated output $\hat{d}_n = \mathbf{w}_n^H \mathbf{x}_n$, where $\mathbf{x}_n = (x_n, x_{n-1}, \ldots, x_{n-N+1})^t$ is the input vector at the nth instant, is as close as possible to the measured output d_n in mean-squared error sense. The affine projection algorithms are iterative procedures to estimate these weights.

The APA class, as mentioned earlier, updates the weights on the basis of multiple input vectors. We use the weight update equation of the NLMS-OCF algorithm [3] for our discussions since it is more general than in the other algorithms of this family (allowing other than unit delay between input vectors) and since the NLMS-OCF update equation is conducive to the analysis that follows. The adaptive filter weights are updated by NLMS-OCF as in

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \mu_0 \mathbf{x}_n + \mu_1 \mathbf{x}_n^1 + \cdots + \mu_M \mathbf{x}_n^M \tag{1}$$

where $M+1$ is the number of input vectors used for adaptation, $\mathbf{x}_n$ is the input vector at the nth instant, $\mathbf{x}_n^k$, for $k = 1, 2, \ldots, M$, is the component of $\mathbf{x}_{n-kD}$ that is orthogonal to $\mathbf{x}_n, \mathbf{x}_{n-D}, \mathbf{x}_{n-2D}, \ldots, \mathbf{x}_{n-(k-1)D}$, D is the delay between input vectors used for adaptation , and μ_k, for $k = 0, 1, \ldots, M$ is chosen as in

$$\mu_k = \begin{cases} \dfrac{\bar{\mu} e_n^*}{\mathbf{x}_n^H \mathbf{x}_n} & \text{for } k = 0, \quad \text{if } \|\mathbf{x}_n\| \neq 0 \\ \dfrac{\bar{\mu} e_n^{k*}}{\mathbf{x}_n^{kH} \mathbf{x}_n^k} & \text{for } k = 1, 2, \ldots, M, \quad \text{if } \|\mathbf{x}_n^k\| \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where

$$\begin{aligned} e_n &= d_n - \mathbf{w}_n^H \mathbf{x}_n, \\ e_n^k &= d_{n-kD} - \mathbf{w}_n^{kH} \mathbf{x}_{n-kD}, \quad \text{for } k = 1, 2, \ldots, M, \quad \text{and} \\ \mathbf{w}_n^k &= \mathbf{w}_n + \mu_0 \mathbf{x}_n + \mu_1 \mathbf{x}_n^1 + \cdots + \mu_{k-1} \mathbf{x}_n^{k-1}. \end{aligned} \tag{3}$$

The constant $\bar{\mu}$ is usually referred to as the step size.

The weight updates generated by APA and GOBA are equivalent to the special case of the weight updates generated by NLMS-OCF, which is shown in (1), with $D = 1$ (see the Appendix). PRA is the special case of APA where the APA weight adaptations are performed once every $M+1$ samples instead of every sample. The flexibility in selecting the vectors used for adaptation, through the choice of D, as provided by NLMS-OCF, has been found to be useful in realizing certain advantageous behavior, such as faster convergence under most conditions and reduction in steady-state error, over the other algorithms in the APA class (which restrict D to be unity) [14]. In the next section, we study the convergence behavior of (1) under certain simplifying assumptions.

III. Convergence Analysis of the Affine Projection Algorithm Class

The convergence analysis is done based on the following assumptions on the signals and the underlying system.

A1) The signal vectors $\{\mathbf{x}_n\}$ have zero mean and are independent and identically distributed (i.i.d.) with covariance matrix

$$\mathbf{R} = E\left[\mathbf{x}_n \mathbf{x}_n^H\right] = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^H \tag{4}$$

where $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$, and $\mathbf{V} = (\boldsymbol{\nu}_1 \ \boldsymbol{\nu}_2 \ \cdots \ \boldsymbol{\nu}_N)$. Here, $\lambda_1, \lambda_2, \ldots, \lambda_N$ are the eigenvalues of $\mathbf{R}$, and $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \ldots, \boldsymbol{\nu}_N$ are the corresponding orthonormal eigenvectors ($\mathbf{V}^H \mathbf{V} = \mathbf{I}$). That is, $\mathbf{V}$ is a unitary matrix.

A2) There exists a true adaptive filter weight $\mathbf{w}^0$ of dimension N such that the corresponding error signal

$$\begin{aligned} e_n &= d_n - \mathbf{w}^{0H} \mathbf{x}_n \\ &\equiv \varepsilon_n \end{aligned} \tag{5}$$

inherits the properties of the measurement noise ε_n, which is a zero mean white noise of variance ξ^0 that is independent of $\{\mathbf{x}_n\}$.

A3) The random vector $\mathbf{x}_n$ is the product of three independent random variables that are i.i.d. That is

$$\mathbf{x}_n = s_n r_n \mathbf{v}_n \tag{6a}$$

where

$$\begin{cases} P\{s_n = \pm 1\} = \dfrac{1}{2} \\ r_n \sim \|\mathbf{x}_n\| \\ P\{\mathbf{v}_n = \boldsymbol{\nu}_i\} = p_i = \dfrac{\lambda_i}{\operatorname{tr}(\mathbf{R})}, \quad i = 1, 2, \ldots, N \end{cases} \tag{6b}$$

where $r_n \sim \|\mathbf{x}_n\|$ means that r_n has the same distribution as the norm of the true input signal vectors.

Assumption A3), which was first introduced by Slock [4], leads to a simple distribution for the vectors $\mathbf{x}_n$ consistent with the actual first- and second-order statistics of the input signal. Assumption A3), as will be seen, makes the convergence analysis tractable. Under assumption A3), the weight update equation of APA can be modified. Since $\mathbf{x}_n$ are either parallel or orthogonal to each other, the orthogonalization step to compute $\mathbf{x}_n^k$, for $k = 1, 2, \ldots, M$, becomes redundant. Hence, (1)–(3)

can be rewritten as shown in (7), (8), shown at the bottom of the next page, and (9), respectively.

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \mu_0\mathbf{x}_n + \mu_1\mathbf{x}_{n-D} + \cdots + \mu_M\mathbf{x}_{n-MD} \quad (7)$$

$$e_n = d_n - \mathbf{w}_n^H\mathbf{x}_n, \quad \text{and}$$

$$e_n^k = d_{n-kD} - \mathbf{w}_n^H\mathbf{x}_{n-kD}, \quad \text{for } k = 1, 2, \ldots, M. \quad (9)$$

[Using (A3), $\mathbf{w}_n^{kH}\mathbf{x}_{n-kD} = (\mathbf{w}_n + \mu_0\mathbf{x}_n + \mu_1\mathbf{x}_{n-D} + \cdots + \mu_{k-1}\mathbf{x}_{n-(k-1)D})^H\mathbf{x}_{n-kD} = \mathbf{w}_n^H\mathbf{x}_{n-kD}$ since $\mathbf{x}_{n-kD} \perp \mathbf{x}_{n-iD}\ \forall i < k$. Hence, (3) can be modified to the form shown in (9).]

To analyze the convergence behavior of (7), first, the weight adaptation is rewritten in terms of the weight error vector $\tilde{\mathbf{w}}_n$, where $\tilde{\mathbf{w}}_n = \mathbf{w}^0 - \mathbf{w}_n$. Using this notation together with (5), we can rewrite e_n^k as $e_n^k = \tilde{\mathbf{w}}_n^H\mathbf{x}_{n-kD} + \varepsilon_{n-kD}$. Combining this result with (7) and (8), the adaptation equation in error form can be obtained as

$$\tilde{\mathbf{w}}_{n+1} = \left[\mathbf{I} - \sum_{j\in J_n} \bar{\mu}\frac{\mathbf{x}_{n-jD}\mathbf{x}_{n-jD}^H}{\mathbf{x}_{n-jD}^H\mathbf{x}_{n-jD}}\right]\tilde{\mathbf{w}}_n - \sum_{l\in J_n} \bar{\mu}\frac{\varepsilon_{n-lD}^*\mathbf{x}_{n-lD}}{\mathbf{x}_{n-lD}^H\mathbf{x}_{n-lD}} \quad (10)$$

where $J_n \subseteq \{0, 1, 2, \ldots, M\}$ is a set of $M+1$ or fewer indices j for which the $\mathbf{x}_{n-jD}$ are orthogonal to each other since $\mu_j = 0$ for $j \notin J_n$. Equation (10) is in a form suitable for convergence analysis. In the absence of noise ε_n, (10) becomes a homogeneous difference equation, whose convergence can be studied. However, with measurement noise, convergence *per se* is not possible; we need to study convergence in the mean and convergence in the mean square. We say that the weights converge in the mean if the expectation of the weight-error vector $\tilde{\mathbf{w}}_n$ approaches zero as the number of iterations n approaches infinity. Convergence in the mean square means that the steady-state value of the covariance $\text{cov}(\tilde{\mathbf{w}}_n)$ of the weight error vector is finite. If these two forms of convergence are satisfied, then the APA algorithm is said to be stable. We begin the convergence analysis with the computation of the weight error vector covariance.

Using (10), the covariance of the weight error vector $\tilde{\mathbf{w}}_n$ is given by

$$\begin{aligned}\text{cov}(\tilde{\mathbf{w}}_{n+1}) = {} & E\left(\left[\mathbf{I} - \sum_{j\in J_n} \bar{\mu}\frac{\mathbf{x}_{n-jD}\mathbf{x}_{n-jD}^H}{\mathbf{x}_{n-jD}^H\mathbf{x}_{n-jD}}\right]\tilde{\mathbf{w}}_n\tilde{\mathbf{w}}_n^H \right.\\ & \left.\times\left[\mathbf{I} - \sum_{l\in J_n} \bar{\mu}\frac{\mathbf{x}_{n-lD}\mathbf{x}_{n-lD}^H}{\mathbf{x}_{n-lD}^H\mathbf{x}_{n-lD}}\right]\right)\\ & + E\left(\left[\sum_{j\in J_n} \bar{\mu}\frac{\varepsilon_{n-jD}^*\mathbf{x}_{n-jD}}{\mathbf{x}_{n-jD}^H\mathbf{x}_{n-jD}}\right]\right.\\ & \left.\times\left[\sum_{l\in J_n} \bar{\mu}\frac{\varepsilon_{n-lD}\mathbf{x}_{n-lD}^H}{\mathbf{x}_{n-lD}^H\mathbf{x}_{n-lD}}\right]\right)\\ & - E\left(\left[\mathbf{I} - \sum_{j\in J_n} \bar{\mu}\frac{\mathbf{x}_{n-jD}\mathbf{x}_{n-jD}^H}{\mathbf{x}_{n-jD}^H\mathbf{x}_{n-jD}}\right]\tilde{\mathbf{w}}_n\right.\\ & \left.\times\left[\sum_{l\in J_n} \bar{\mu}\frac{\varepsilon_{n-lD}\mathbf{x}_{n-lD}^H}{\mathbf{x}_{n-lD}^H\mathbf{x}_{n-lD}}\right]\right)\\ & - E\left(\left[\sum_{j\in J_n} \bar{\mu}\frac{\varepsilon_{n-jD}^*\mathbf{x}_{n-jD}}{\mathbf{x}_{n-jD}^H\mathbf{x}_{n-jD}}\right]\tilde{\mathbf{w}}_n^H\right.\\ & \left.\times\left[\mathbf{I} - \sum_{l\in J_n} \bar{\mu}\frac{\mathbf{x}_{n-lD}\mathbf{x}_{n-lD}^H}{\mathbf{x}_{n-lD}^H\mathbf{x}_{n-lD}}\right]\right). \end{aligned} \quad (11)$$

If the dependency of $\tilde{\mathbf{w}}_n$ on past measurement noise is neglected, using that ε_n is of zero mean, the last two terms of the above expression vanish. Furthermore, if we neglect[1] the dependency of $\tilde{\mathbf{w}}_n$ on the past input vectors that appear in the first term of the above expression and use A2) to simplify the second term, we can rewrite (11) as

$$\begin{aligned}\text{cov}(\tilde{\mathbf{w}}_{n+1}) = {} & E\left(\left[\mathbf{I} - \sum_{j\in J_n} \bar{\mu}\frac{\mathbf{x}_{n-jD}\mathbf{x}_{n-jD}^H}{\mathbf{x}_{n-jD}^H\mathbf{x}_{n-jD}}\right]\right.\\ & \left.\times\,\text{cov}(\tilde{\mathbf{w}}_n)\left[\mathbf{I} - \sum_{l\in J_n} \bar{\mu}\frac{\mathbf{x}_{n-lD}\mathbf{x}_{n-lD}^H}{\mathbf{x}_{n-lD}^H\mathbf{x}_{n-lD}}\right]\right)\\ & + E\left(\bar{\mu}^2\sum_{j\in J_n}|\varepsilon_{n-jD}|^2\frac{\mathbf{x}_{n-jD}\mathbf{x}_{n-jD}^H}{\|\mathbf{x}_{n-jD}\|^2\mathbf{x}_{n-jD}^H\mathbf{x}_{n-jD}}\right). \end{aligned} \quad (12)$$

Using A3), we can rewrite the outer- to inner-product ratios as

$$\begin{aligned}\frac{\mathbf{x}_{n-jD}\mathbf{x}_{n-jD}^H}{\mathbf{x}_{n-jD}^H\mathbf{x}_{n-jD}} &= \frac{s_{n-jD}r_{n-jD}\mathbf{v}_{n-jD}\mathbf{v}_{n-jD}^H r_{n-jD}s_{n-jD}}{s_{n-jD}^2 r_{n-jD}^2\|\mathbf{v}_{n-jD}\|^2}\\ &= \mathbf{v}_{n-jD}\mathbf{v}_{n-jD}^H \end{aligned} \quad (13)$$

[1]In the case of PRA, no approximation is involved in this step since $\bar{\mathbf{w}}_n$ is independent of the input vectors used for adaptation.

$$\mu_k = \begin{cases} \dfrac{\bar{\mu}e_n^*}{\mathbf{x}_n^H\mathbf{x}_n}, & \text{for } k = 0, \quad \text{if } \|\mathbf{x}_n\| \neq 0 \\ \dfrac{\bar{\mu}e_n^{k*}}{\mathbf{x}_{n-kD}^H\mathbf{x}_{n-kD}}, & \text{for } k = 1, 2, \ldots, M, \quad \text{if } \mathbf{x}_{n-kD} \perp \mathbf{x}_{n-iD} \quad \forall i < k \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

where $\mathbf{v}_{n-jD}$ is one of $\{\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \ldots, \boldsymbol{\nu}_N\}$. Note that the above result is independent of the norm of $\mathbf{x}_{n-jD}$. Now, substituting (13) into (12) we get

$$\begin{aligned}\text{cov}(\tilde{\mathbf{w}}_{n+1}) &= E\left(\left[\mathbf{I} - \sum_{j\in J_n} \bar{\mu}\mathbf{v}_{n-jD}\mathbf{v}_{n-jD}^H\right]\right. \\ &\quad \left.\times\ \text{cov}(\tilde{\mathbf{w}}_n)\left[\mathbf{I} - \sum_{l\in J_n} \bar{\mu}\mathbf{v}_{l-jD}\mathbf{v}_{l-jD}^H\right]\right) \\ &\quad + E\left(\bar{\mu}^2 \sum_{j\in J_n} |\varepsilon_{n-jD}|^2 \frac{1}{r_{n-jD}^2}\mathbf{v}_{n-jD}\mathbf{v}_{n-jD}^H\right). \quad (14)\end{aligned}$$

Since ε_n is independent of $\mathbf{x}_n$ and r_n is independent of $\mathbf{v}_n$, from A2) and A3), respectively, we can rewrite (14) as

$$\begin{aligned}\text{cov}(\tilde{\mathbf{w}}_{n+1}) &= E\left(\left[\mathbf{I} - \sum_{k\in K_n} \bar{\mu}\boldsymbol{\nu}_k\boldsymbol{\nu}_k^H\right]\text{cov}(\tilde{\mathbf{w}}_n)\left[\mathbf{I} - \sum_{l\in K_n} \bar{\mu}\boldsymbol{\nu}_k\boldsymbol{\nu}_k^H\right]\right) \\ &\quad + \bar{\mu}^2\xi_0 E\left(\frac{1}{r^2}\right) E\left(\sum_{k\in K_n} \boldsymbol{\nu}_k\boldsymbol{\nu}_k^H\right) \quad (15)\end{aligned}$$

where

$$\begin{aligned}K_n &= \left\{k : \exists j \in J_n \ni \frac{\mathbf{x}_{n-jD}\mathbf{x}_{n-jD}^H}{\mathbf{x}_{n-jD}^H\mathbf{x}_{n-jD}} = \boldsymbol{\nu}_k\boldsymbol{\nu}_k^H\right\} \\ &\subseteq \{1, 2, \ldots, N\}. \quad (16)\end{aligned}$$

Let us define the diagonal elements of the transformed covariance matrix $\mathbf{V}^H\text{cov}(\tilde{\mathbf{w}}_n)\mathbf{V}$ as $\tilde{\lambda}_{n,i}$ for $i = 1, 2, \ldots, N$. That is

$$[\mathbf{V}^H\text{cov}(\tilde{\mathbf{w}}_n)\mathbf{V}]_{ii} = \boldsymbol{\nu}_i^H\text{cov}(\tilde{\mathbf{w}}_n)\boldsymbol{\nu}_i = \tilde{\lambda}_{n,i}. \quad (17)$$

Note that this does not mean that $\mathbf{V}^H\text{cov}(\tilde{\mathbf{w}}_n)\mathbf{V}$ is a diagonal matrix.

With the above notation, the pre- and post-multiplication of (15) by and $\boldsymbol{\nu}_i^H$ and $\boldsymbol{\nu}_i$, respectively, results in

$$\begin{aligned}\tilde{\lambda}_{n+1,i} &= E\left(\boldsymbol{\nu}_i^H\left[\mathbf{I} - \sum_{k\in K_n} \bar{\mu}\boldsymbol{\nu}_k\boldsymbol{\nu}_k^H\right]\right. \\ &\quad \left.\times\ \text{cov}(\tilde{\mathbf{w}}_n)\left[\mathbf{I} - \sum_{k\in K_n} \bar{\mu}\boldsymbol{\nu}_k\boldsymbol{\nu}_k^H\right]\boldsymbol{\nu}_i\right) \\ &\quad + \bar{\mu}^2\xi^0 E\left(\frac{1}{r^2}\right) E\left(\boldsymbol{\nu}_i^H\left[\sum_{k\in K_n} \boldsymbol{\nu}_k\boldsymbol{\nu}_k^H\right]\boldsymbol{\nu}_i\right). \quad (18)\end{aligned}$$

From the orthonormality of the $\boldsymbol{\nu}_k$'s,

$$\boldsymbol{\nu}_i^H \sum_{k\in K_n} \boldsymbol{\nu}_k\boldsymbol{\nu}_k^H = \begin{cases}\boldsymbol{\nu}_i^H, & \text{if } i \in K_n \\ \mathbf{0}, & \text{if } i \notin K_n.\end{cases} \quad (19)$$

Using the above result, (18) can be rewritten as

$$\begin{aligned}\tilde{\lambda}_{n+1,i} &= \boldsymbol{\nu}_i^H\text{cov}(\tilde{\mathbf{w}}_n)\boldsymbol{\nu}_i + E\left(\boldsymbol{\nu}_i^H\left[\sum_{k\in K_n} \bar{\mu}\boldsymbol{\nu}_k\boldsymbol{\nu}_k^H\right]\right. \\ &\quad \left.\times\ \text{cov}(\tilde{\mathbf{w}}_n)\left[\sum_{l\in K_n} \bar{\mu}\boldsymbol{\nu}_l\boldsymbol{\nu}_l^H\right]\boldsymbol{\nu}_i\right) \\ &\quad - E\left(\boldsymbol{\nu}_i^H\text{cov}(\tilde{\mathbf{w}}_n)\left[\sum_{l\in K_n} \bar{\mu}\boldsymbol{\nu}_l\boldsymbol{\nu}_l^H\right]\boldsymbol{\nu}_i\right) \\ &\quad - E\left(\boldsymbol{\nu}_i^H\left[\sum_{k\in K_n} \bar{\mu}\boldsymbol{\nu}_k\boldsymbol{\nu}_k^H\right]\text{cov}(\tilde{\mathbf{w}}_n)\boldsymbol{\nu}_i\right) \\ &\quad + \bar{\mu}^2\xi^0 E\left(\frac{1}{r^2}\right) E\left(\boldsymbol{\nu}_i^H\left[\sum_{k\in K_n} \boldsymbol{\nu}_k\boldsymbol{\nu}_k^H\right]\boldsymbol{\nu}_i\right) \\ &= \tilde{\lambda}_{n,i}[1 - \bar{\mu}(2 - \bar{\mu})P(i \in K_n)] \\ &\quad + \bar{\mu}^2\xi^0 E\left(\frac{1}{r^2}\right) P(i \in K_n). \quad (20)\end{aligned}$$

The probability $P(i \in K_n)$ is the same as the probability of drawing (with replacement) the ball marked i, at least once in $M+1$ trials, from a collection of N balls marked $1, 2, \ldots, N$, where the probability of drawing the ball marked j is p_j. Hence

$$P(i \in K_n) = 1 - (1 - p_i)^{M+1}. \quad (21)$$

By substituting (21) into (20), we get

$$\tilde{\lambda}_{n+1,i} = (1 - \alpha\beta_i)\tilde{\lambda}_{n,i} + \bar{\mu}^2\xi^0 E\left(\frac{1}{r^2}\right)\beta_i \quad (22)$$

where $\alpha = \bar{\mu}(2 - \bar{\mu})$, and $\beta_i = 1 - (1 - p_i)^{M+1}$.

The following observations can be made from (22).

Observation 1: $0 < \bar{\mu} < 2$ *is a necessary and sufficient condition for the APA class to be stable*. Let us first look at the mean-squared convergence. The error e_n in the output estimate is given by

$$e_n = \tilde{\mathbf{w}}_n^H\mathbf{x}_n + \varepsilon_n. \quad (23)$$

Using A2), the mean-squared error $\xi_n = E(e_n e_n^*)$ in the output estimate can be written as

$$\begin{aligned}\xi_n &= \xi^0 + E\left(\|\tilde{\mathbf{w}}_n^H\mathbf{x}_n\|^2\right) \\ &= \xi^0 + \text{tr}[\mathbf{R}\,\text{cov}(\tilde{\mathbf{w}}_n)] \\ &= \xi^0 + \text{tr}[\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^H\text{cov}(\tilde{\mathbf{w}}_n)] \\ &= \xi^0 + \text{tr}[\boldsymbol{\Lambda}\mathbf{V}^H\text{cov}(\tilde{\mathbf{w}}_n)\mathbf{V}] \\ &= \xi^0 + \sum_{i=1}^{N} \lambda_i\tilde{\lambda}_{n,i}. \quad (24)\end{aligned}$$

From (24), we see that the mean-squared error converges if $\tilde{\lambda}_{n,i}$ converges. If $\bar{\mu} \in (0, 2)$ and the input signal is sufficiently rich ($p_i \neq 0$ for any i), then $\alpha \in (0, 1]$, and $0 \leq (1 - \alpha\beta_i) < 1$; this guarantees the convergence of $\tilde{\lambda}_{n,i}$ in (22). If $\bar{\mu} \notin (0, 2)$, then $\alpha \leq 0$, and $(1 - \alpha\beta_i) \geq 1$; hence, $\tilde{\lambda}_{n,i}$ does not converge.

Thus, provided $\bar{\mu} \in (0,2)$ and the input is sufficiently rich, the steady-state solution of (22) is given by

$$\lim_{n\to\infty} \tilde{\lambda}_{n,i} = \frac{\bar{\mu}}{2-\bar{\mu}} \xi^0 E\left(\frac{1}{r^2}\right). \tag{25}$$

Combining (24) and (25), the steady-state (final) mean-squared error is given by

$$\xi_\infty = \xi^0 \left[1 + \frac{\bar{\mu}}{2-\bar{\mu}} E\left(\frac{1}{r^2}\right) \mathrm{tr}(\mathbf{R})\right] < \infty. \tag{26}$$

Using (24), the finiteness of the steady-state mean-squared error implies the finiteness of $\mathrm{cov}(\tilde{\mathbf{w}}_n)$ in steady state. That is $\mathrm{cov}(\tilde{\mathbf{w}}_n)$ is asymptotically stable. Thus, for sufficiently rich inputs, $\bar{\mu} \in (0,2)$ is a necessary and sufficient condition for convergence in mean square.

Now, we analyze the convergence in the mean. After we neglect the dependence of $\tilde{\mathbf{w}}_n$ on the past input vectors, taking expectation on both sides of (10) results in

$$E(\tilde{\mathbf{w}}_{n+1}) = E(\tilde{\mathbf{w}}_n) - E\left(\bar{\mu} \sum_{k\in K_n} \boldsymbol{\nu}_k \boldsymbol{\nu}_k^H \tilde{\mathbf{w}}_n\right). \tag{27}$$

Here, we used (16) to replace the outer- to inner-product ratios with $\boldsymbol{\nu}_k \boldsymbol{\nu}_k^H$ and used A2) to conclude that the expected value of the term with ε_n vanishes.

Define vector $\boldsymbol{\rho}_n$ as the representation of $E(\tilde{\mathbf{w}}_n)$ in terms of the orthonormal vectors $\{\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \ldots, \boldsymbol{\nu}_N\}$. That is

$$\boldsymbol{\rho}_n \equiv \mathbf{V}^H E(\tilde{\mathbf{w}}_n). \tag{28}$$

Therefore

$$\rho_{n,i} = \boldsymbol{\nu}_i^H E(\tilde{\mathbf{w}}_n) = E(\boldsymbol{\nu}_i^H \tilde{\mathbf{w}}_n). \tag{29}$$

Using this notation, premultiplication of (27) by $\mathbf{v}_i^H$ results in

$$\rho_{n+1,i} = \rho_{n,i} - E\left(\bar{\mu} \boldsymbol{\nu}_i^H \sum_{k\in K_n} \boldsymbol{\nu}_k \boldsymbol{\nu}_k^H \tilde{\mathbf{w}}_n\right). \tag{30}$$

Using (19) and (21), (30) can be rewritten as

$$\rho_{n+1,i} = (1 - \bar{\mu}\beta_i)\rho_{n,i}. \tag{31}$$

From (31), we see that $\rho_{n,i}$ converges to zero if and only if $|1 - \bar{\mu}\beta_i| < 1$. For sufficiently rich inputs, we have $0 < \beta_i \leq 1$. Hence, $\bar{\mu} \in (0,2)$ is a sufficient condition for $\rho_{n,i}$ to converge. Consequently, if $\bar{\mu} \in (0,2)$, $\boldsymbol{\rho}_n$ converges to zero exponentially as n approaches infinity. Since $\{\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \ldots, \boldsymbol{\nu}_N\}$ forms an orthonormal basis, $\|E(\tilde{\mathbf{w}}_n)\| = \|\boldsymbol{\rho}_n\|$. Hence, $E(\tilde{\mathbf{w}}_n)$ converges to zero as n approaches infinity. In other words, APA is an asymptotically unbiased estimator of the weights. Thus, $\bar{\mu} \in (0,2)$ is a sufficient condition for convergence in mean. Combining the conditions for mean and mean-squared convergence, $0 < \bar{\mu} < 2$ is a necessary and sufficient condition for the APA class to be stable. Earlier, this algorithm stability condition was made plausible geometrically for the noiseless case [3], [7].

Observation 2: The convergence behavior of the mean-squared error ξ_n for the noiseless case, viz. $\xi^0 = 0$, is exponential, as given in (37). We begin the analysis by making a few assumptions on initial conditions. Assume that no *a priori* information on the system is available and, hence, that the typical initial estimate for the weights $\mathbf{w}_0 = \mathbf{0}$ is used. We use the maximum entropy assumption for the optimal weights [4]. That is, $\mathbf{w}^0$ has equal components along all eigenvectors of $\mathbf{R}$. For example

$$\mathbf{w}^0 = \sqrt{\frac{\sigma_d^2 - \xi^0}{\mathrm{tr}(\mathbf{R})}} \mathbf{V} 1_N \tag{32}$$

where σ_d^2 is the variance of the output signal d_n, and $1_N \equiv [1\ 1\ \cdots\ 1]^T$ satisfies the maximum entropy assumption. For these values of the optimal weight $\mathbf{w}^0$ and the initial estimate $\mathbf{w}_0$, assuming $\xi^0 = 0$

$$\mathrm{cov}(\tilde{\mathbf{w}}_0) = E(\tilde{\mathbf{w}}_0 \tilde{\mathbf{w}}_0^H) = \mathbf{w}^0 \mathbf{w}^{0H} = \frac{\sigma_d^2}{\mathrm{tr}(\mathbf{R})} \mathbf{V} 1_N 1_N^H \mathbf{V}^H. \tag{33}$$

Using the fact that $\mathbf{V}$ is unitary, it follows that

$$\mathbf{V}^H \mathrm{cov}(\tilde{\mathbf{w}}_0) \mathbf{V} = \frac{\sigma_d^2}{\mathrm{tr}(\mathbf{R})} 1_N 1_N^H. \tag{34}$$

The above is a matrix with $\sigma_d^2/\mathrm{tr}(\mathbf{R})$ as all its entries. Hence, using (17), we get

$$\tilde{\lambda}_{0,i} = \sigma_d^2/\mathrm{tr}(\mathbf{R}) \quad \forall i. \tag{35}$$

Solving (22), using (35) as the initial condition, and substituting the solution in (24), we get the mean-squared error as

$$\xi_n = \sum_{i=1}^{N} \lambda_i (1 - \alpha\beta_i)^n \frac{\sigma_d^2}{\mathrm{tr}(\mathbf{R})}. \tag{36}$$

From A3), $\lambda_i/\mathrm{tr}(\mathbf{R}) = p_i$, so that we can rewrite (36) as

$$\xi_n = \sigma_d^2 \sum_{i=1}^{N} (1 - \alpha\beta_i)^n p_i. \tag{37}$$

Hence, (37) describes the theoretical convergence behavior of the APA class of algorithms under noise-free conditions.

Observation 3: APA converges faster than NLMS; as more input vectors are used, the convergence rate itself improves, whereas the rate of this improvement decreases. From (22), we see that the rate of convergence depends on the factor $(1-\alpha\beta_i)$, where $\alpha = \bar{\mu}(2 - \bar{\mu}) \leq 1$, and $\beta_i = 1 - (1 - p_i)^{M+1} \leq 1$. Note that the values of α, and, hence, the convergence rates, are the same for step sizes $\bar{\mu}$ and $2 - \bar{\mu}$ for $\bar{\mu} \in (0,2)$. However, as will be shown in Observation 5, the steady-state mean-squared error increases as $\bar{\mu}$ increases. In view of this, it is better to use a step size $\bar{\mu} \in (0,1]$. As we can see from (22), faster convergence occurs for values of $(1-\alpha\beta_i)$ closer to 0 (equivalently α and β_i closer to 1). Hence, we want $\alpha = 1$ for fast convergence. Equivalently, $\bar{\mu} = 1$ is the optimum step size value for fastest convergence. Furthermore, increasing the number of input vectors $(M+1)$ used for adaptation increases the convergence rate since, as $(M+1)$ increases, β_i gets closer to 1. This explains the faster convergence of APA over NLMS. Fig. 2 shows a plot of the convergence rate factor $(1 - \alpha\beta_i)$ for different values of M and different values of p_i, with $\bar{\mu} = 1$. It is evident from this plot that the convergence rate factor has an exponential dependence on M. That is $(1 - \alpha\beta_i)$ behaves like η_i^M for some $\eta_i < 1$. Hence, for large enough values of n, with $\bar{p}_i$ denoting

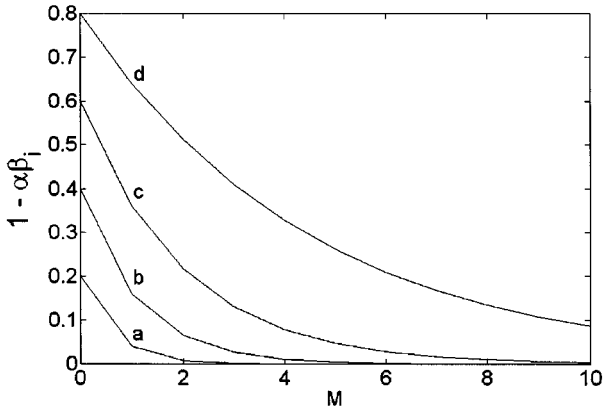


Fig. 2. Dependence of convergence rate factor $(1-\alpha\beta_i)$ on M (a) $p_i = 0.2$. (b) $p_i = 0.4$. (c) $p_i = 0.6$. (d) $p_i = 0.8$.

the total probability mass associated with the largest of the η_i, (37) can be approximated as

$$\xi_n \approx \sigma_d^2 \left(\max_i \eta_i\right)^{Mn} \bar{p}_i. \tag{38}$$

Equivalently

$$\xi_{n,\mathrm{dB}} \approx 10\log_{10}\sigma_d^2\bar{p}_i + 10Mn\log_{10}\left(\max_i \eta_i\right). \tag{39}$$

Thus, for large enough n, the slope of the learning curve (plot showing mean-squared error in decibels versus iteration number) depends linearly on M. If we next define the time to (reach) steady state T_{SS} as a performance index of the algorithm, the rate at which the performance improves diminishes as M increases. This explains the phenomenon that Gay and Tavathia observed in their simulation results [8].

Observation 4: If the input is white, the learning curve is linear, and the mean-squared error drops by 20 dB in about $5N/(M+1)$ iterations. Assume that the input $\mathbf{x}_n$ to the adaptive filter is white. In this case, all the p_i's are equal. That is

$$p_i = \frac{1}{N} \quad \text{for } i = 1, 2, \ldots, N. \tag{40}$$

Therefore, if the step size is chosen to be unity, the convergence rate factor for white noise can be written as

$$(1-\alpha\beta_i) = \left(1 - \frac{1}{N}\right)^{M+1}. \tag{41}$$

Substituting (41) into (37), the mean-squared error convergence is given by

$$\begin{aligned}\xi_n &= \sigma_d^2 \sum_{i=1}^{N} \frac{1}{N}\left(1-\frac{1}{N}\right)^{n(M+1)} \\ &= \sigma_d^2 \left(1-\frac{1}{N}\right)^{n(M+1)}.\end{aligned} \tag{42}$$

Hence, the mean-squared error in decibels can be written as

$$\begin{aligned}\xi_{n,\mathrm{dB}} &= 10\log_{10}\sigma_d^2 + 10n(M+1)\log_{10}\left(1-\frac{1}{N}\right) \\ &= 10\log_{10}\sigma_d^2 - 4.343\frac{n(M+1)}{N}.\end{aligned} \tag{43}$$

Thus, the learning curve for a white input is linear and the mean squared error drops by about 20 dB in $5N/(M+1)$ iterations for $\bar{\mu} = 1$. This also means that longer filters exhibit slower convergence. This observation also corroborates the idea that the convergence rate can be improved by starting with a smaller number of taps in the adaptive filter and then gradually increasing the number of taps until the desired order is reached. A similar idea was exploited to accelerate the convergence of LMS [6].

Observation 5: The misadjustment of the APA class is independent of M. Using (26), the misadjustment, which is defined as the ratio of excess mean-squared error to minimum mean-squared error, equals

$$\mathcal{M} = \frac{\xi_\infty - \xi^0}{\xi^0} = \frac{\bar{\mu}}{2-\bar{\mu}} E\left(\frac{1}{r^2}\right)\mathrm{tr}(\mathbf{R}). \tag{44}$$

Note the independence of (44) of M. In fact, it is the same as the misadjustment of the NLMS algorithm (NLMS is the special case of APA with $M = 0$) with the same $\bar{\mu}$. The independence of (44) of M is, perhaps, due to the fact that we neglected dependence of $\tilde{\mathbf{w}}_n$ on past measurement noise while going from (11) to (12). Simulation results indicate a "weak" dependence of misadjustment on M. As shown in Observation 3, the convergence rate improves with increasing M. Thus, APA provides a way to increase the convergence rate without compromising too much on misadjustment and, hence, the steady state mean-squared error of APA. This is yet another advantage, so far unreported, of APA over NLMS.

Observation 6: NLMS is the special case of APA with $M = 0$. If $M = 0$, then $\beta_i = p_i$, and difference equation (22), which describes the behavior of $\tilde{\lambda}_{n,i}$, becomes

$$\tilde{\lambda}_{n+1,i} = (1-\alpha p_i)\tilde{\lambda}_{n,i} + \bar{\mu}^2\xi^0 E\left(\frac{1}{r^2}\right)p_i. \tag{45}$$

Similarly, the NLMS mean-squared error convergence behavior is given by

$$\xi_n = \sigma_d^2 \sum_{i=1}^{N} (1-\alpha p_i)^n p_i. \tag{46}$$

These results match the earlier results derived for NLMS under the same assumptions [4]. From Observation 4, the learning curve of NLMS drops by 20 dB in about $5N$ iterations for $\bar{\mu} = 1$. This result conforms to Rupp's observation on the convergence speed of NLMS [11].

A Special Comment for PRA

The PRA attempts to reduce the complexity of APA by adapting the weights once every $M+1$ samples instead of every sample. Hence, the analysis above gives *mutatis mutandis* the results for PRA. The diagonal elements of the transformed covariance matrix of the weight estimation error, which is defined in (17), become, for PRA

$$\tilde{\lambda}_{n+1,i} = \begin{cases} \tilde{\lambda}_{n,i}, & \text{if } ((n+1))_{M+1} \neq 0 \\ (1-\alpha\beta_i)\tilde{\lambda}_{n,i} + \bar{\mu}^2\xi^0 E\left(\frac{1}{r^2}\right)\beta_i, & \text{if } ((n+1))_{M+1} = 0 \end{cases} \tag{47}$$

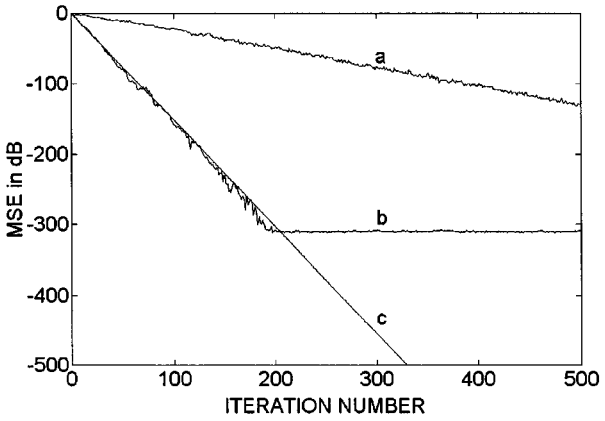


Fig. 3. Learning curves of APA for white input using $\bar{\mu} = 1.0$ (a) Simulated with $D = 1$. (b) Simulated with $D = 32$. (c) Theoretical. (Input: White noise. System: FIR(31), $\xi^0 = 0$, and $M = 10$).

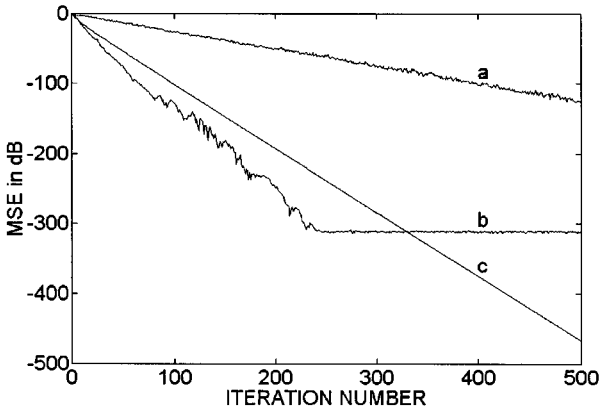


Fig. 4. Learning curves of APA for reasonably colored input using $\bar{\mu} = 1.0$ (a) Simulated with $D = 1$. (b) Simulated with $D = 32$. (c) Theoretical. (Input: AR(1), pole at 0.25. System: FIR(31), $\xi^0 = 0$, and $M = 10$).

where $((n))_M$ denotes n modulo M. The mean-squared error ξ_n of PRA is thus given by

$$\xi_n = \sigma_d^2 \sum_{i=1}^{N} (1 - \alpha\beta_i)^{\lfloor \frac{n}{M+1} \rfloor} p_i \tag{48}$$

where $\lfloor x \rfloor$ denotes the largest integer that is less than or equal to x.

IV. Verification Using Simulation

In this section, we demonstrate the validity of the analytical results presented in Section III and discuss limitations introduced by the assumptions. Simulation and theoretical results corresponding to three different types of signals, viz. white, reasonably colored, and highly colored, are shown. The reasonably and highly colored signals are generated as a Gaussian first-order autoregressive process with a pole at 0.25 and 0.95, respectively. The system to be identified has a 32-point long impulse response computed according to (32) for each case, and hence, the impulse response satisfies the maximum entropy assumption. The delay line of the adaptive filter is initialized with true data values (soft initialization) in all simulations, and $\mathbf{w}_0 = \mathbf{0}$ is used as the initial estimate for the weights. The measurement noise is assumed to be absent ($\xi^0 = 0$) unless noted otherwise. The simulation results shown are obtained by ensemble averaging over 100 independent trials of the experiment.

Fig. 3 shows the results obtained using a white input signal. The weight updates are performed with 11 input vectors, i.e., $M = 10$. The steady-state MSE is limited in simulation to around -325 dB because of the quantization errors introduced in the calculations. We see that the theoretical result, as given by (38), is very close to the simulated result when $D = 32$ and that there is an appreciable deviation between the theoretical and simulated results when $D = 1$. This is because of the independence assumption that we used in the analysis. The input vectors used for a particular weight update are truly independent when $D = 32$, whereas this is not true when $D = 1$. This is an advantage of NLMS-OCF, which allows $D > 1$.

The results obtained using the reasonably colored signal as input are shown in Fig. 4. The simulation result is closer to the theoretical result when $D = 32$ than when $D = 1$ since the input vectors used for weight updates are more nearly independent when $D = 32$ than when $D = 1$.

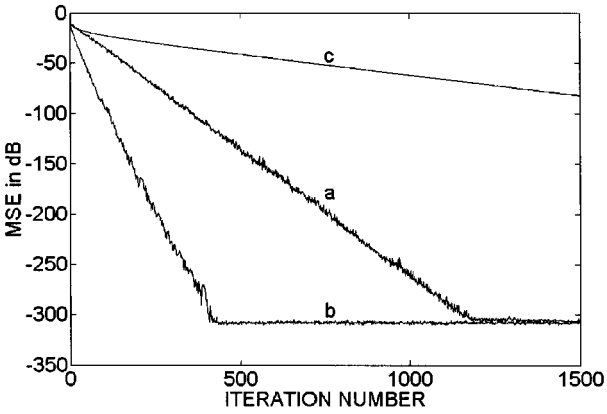


Fig. 5. Learning curves of APA for highly colored input using $\bar{\mu} = 1.0$ (a) Simulated with $D = 1$. (b) Simulated with $D = 32$. (c) Theoretical. (Input: AR(1), pole at 0.95. System: FIR(31), $\xi^0 = 0$, and $M = 10$).

Results, for the highly colored signal as input, which are similar to the results shown in Figs. 3 and 4, are shown in Fig. 5. We see that there is a larger deviation between the theoretical and simulation results in this case than in the white noise and reasonably colored case. We would expect this behavior since the highly correlated input violates the independence assumption more strongly than the other two inputs.

From Figs. 3–5, we note that the convergence for the $D = 1$ case does not depend on the color of the input signal; curve (a) reaches -130 dB at iteration 500. For the $D = 32$ case, convergence is faster than for $D = 1$, with dependence on the color of the input for the highly colored input causing some slowing down in convergence.

The independence assumption of the input vectors is used to claim that the weight estimate $\mathbf{w}_n$ is independent of the input vectors $\mathbf{x}_k$ for all $k \leq n$. The dependence of $\mathbf{w}_n$ on the past input vectors can also be reduced by using a smaller value for the step size. For this reason, we expect the simulation results to be in better agreement with the theoretical results for smaller step-size values. This, in fact, is true, as can be seen from comparing the results in Figs. 4 and 6, which are obtained using

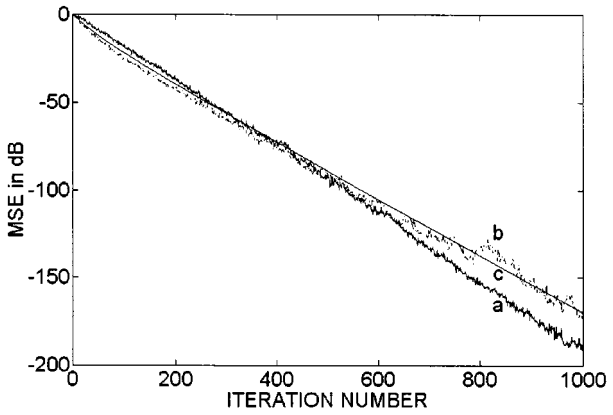


Fig. 6. Learning curves of APA for reasonably colored input using $\bar{\mu} = 0.1$ (a) Simulated with $D = 1$. (b) Simulated with $D = 32$. (c) Theoretical. (Input: AR(1), pole at 0.25. System: FIR(31), $\xi^0 = 0$, and $M = 10$).

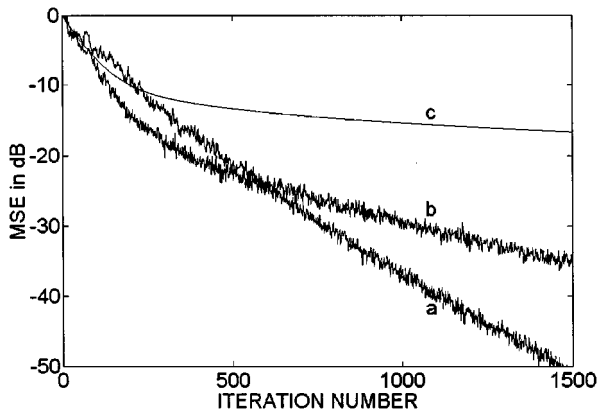


Fig. 7. Learning curves of APA for highly colored input using $\bar{\mu} = 0.01$ (a) Simulated with $D = 1$. (b) Simulated with $D = 32$. (b) Theoretical. (Input: AR(1), pole at 0.95. System: FIR(31), $\xi^0 = 0$, and $M = 10$).

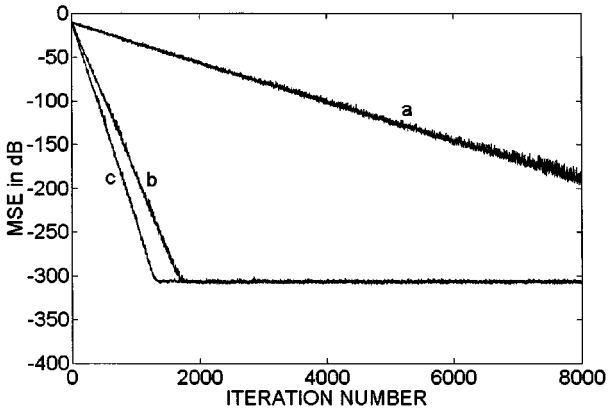


Fig. 8. Simulated learning curves of APA for highly colored input—Various M (a) $M = 0$ (NLMS). (b) $M = 2$. (c) $M = 8$. (Input: AR(1), pole at 0.95. System: FIR(31), $\xi^0 = 0$, $\bar{\mu} = 1.0$, and $D = 1$).

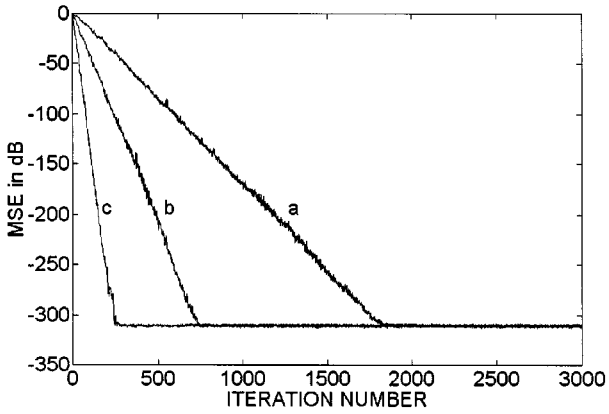


Fig. 9. Simulated learning curves of APA for white input—Various M (a) $M = 0$ (NLMS). (b) $M = 2$. (c) $M = 8$. (Input: White noise. System: FIR(31), $\xi^0 = 0$, $\bar{\mu} = 1.0$, and $D = 32$).

the reasonably colored signal. For an identical value of D, input signal, and system, the theoretical result is matched better by the simulation result when $\bar{\mu} = 0.1$ than when $\bar{\mu} = 1$. In addition, note that the convergence rate is slower with $\bar{\mu} = 0.1$ than with $\bar{\mu} = 1$.

The simulation results and theoretical results for the highly colored input signal are shown in Fig. 7. Here, in addition, the simulation result with $D = 32$ is closer to the theoretical result than the simulation result with $D = 1$. We see that there is a large deviation between the theoretical and simulation results in this case (even with a small value of $\bar{\mu}$). This is again due to the strong dependency between input vectors used for successive adaptations. Hence, the weight estimate $\mathbf{w}_n$ is not really independent of the input vectors $\mathbf{x}_k$. Note in this case, where $\bar{\mu}$ is small, that eventually, the convergence rate for $D = 1$ exceeds that for $D = 32$. Recall that for fast convergence, $\bar{\mu} = 1.0$ is optimal and that in Figs. 3–5, the convergence for $D = 32$ is faster than for $D = 1$. The latter behavior is not universal, as the results in Fig. 7 illustrate.

Fig. 8 shows the simulation results obtained by using a different number of vectors $(M+1)$ for adaptation. The highly colored signal is used as the input. While for $M = 0$ the steady state is projected to be reached in about 14 000 iterations, the steady state is reached for $M = 2$ and 8 in about 1600 and 1200 iterations, respectively. Thus, the improvement in time-to-steady-state T_{SS} achieved by increasing M from 2 to 8 is less than the improvement achieved by increasing M from 0 to 2. This confirms Observation 3 from the analytical results—the T_{SS} improvement rate diminishes as M increases. It is worthwhile to point out that the characteristic predicted by our analysis holds, even though the highly colored input signal does not conform to our assumptions on the data.

The simulation results with white noise input, for different values of M, as shown in Fig. 9, corroborate Observation 4. Although the theoretical predictions for the slope of the learning curves for $M = 0$, 2, and 8, using (42), are 0.14, 0.41, and 1.2 dB/iteration, respectively, the corresponding slopes estimated from the simulation results are about 0.17, 0.42, and 1.3 dB/iteration respectively. It is interesting to note that APA provides an improvement in convergence rate not only for colored input but also for white input. Even when the delay is chosen to be unity, with white input, the convergence rate of APA improves as the number of vectors used for adaptation increases. This shows that APA is not merely a decorrelating algorithm since the decorrelating-algorithm interpretation [11] suggests that APA will not converge faster than NLMS when the input is white, which cannot be decorrelated any further by APA.

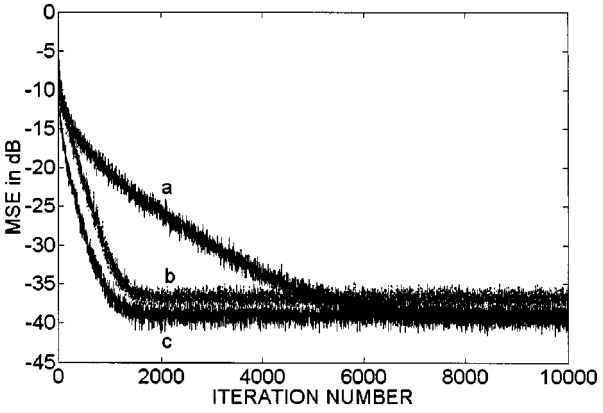


Fig. 10. Simulated learning curves of APA—Misadjustment/convergence rate tradeoff (a) $M = 0$ (NLMS) and $\bar{\mu} = 0.25$. (b) $M = 0$ (NLMS) and $\bar{\mu} = 1.0$. (c) $M = 2$ and $\bar{\mu} = 0.25$. (Input: AR(1), pole at 0.95. System: FIR(31), $\xi^0 = 10^{-4}$, and $D = 32$).

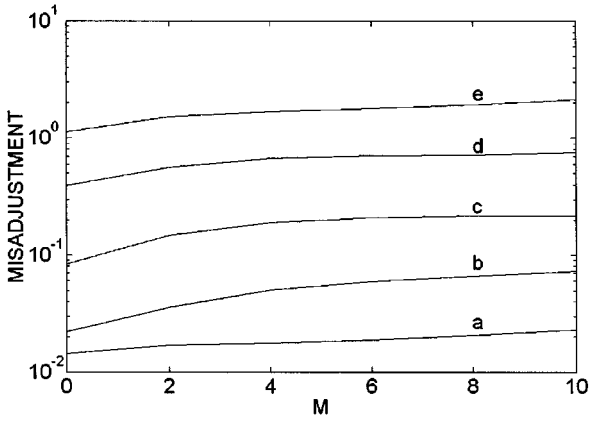


Fig. 11. Dependence of Misadjustment on step size (a) $\bar{\mu} = 0.001$. (b) $\bar{\mu} = 0.01$. (c) $\bar{\mu} = 0.1$. (d) $\bar{\mu} = 0.5$. (e) $\bar{\mu} = 1.0$. (Input: AR(1), pole at 0.95. System: FIR(31), $\xi^0 = 10^{-4}$, and $D = 32$).

Observation 5 suggested that APA provides a way to improve the convergence rate without compromising on misadjustment. The following experiment corroborates this observation. Fig. 10(a) shows the learning curve of NLMS with a step size $\bar{\mu}$ of 0.25. We see that the algorithm takes about 8000 iterations to converge. The misadjustment $\mathcal{M}$ is 0.2062 for this case. An improvement in convergence can be achieved either by using a larger value of step size $\bar{\mu}$ or by using the affine projection algorithm (that is, by using more input vectors for the weight update). Figs. 10(b) and (c) show the learning curves obtained by using NLMS with $\bar{\mu} = 1$ and by using APA with $M = 2$ (and $\bar{\mu} = 0.25$), respectively. In both these cases, we see faster convergence than for NLMS with $\bar{\mu} = 0.25$. It is evident that their individual convergence rates are nearly comparable, whereas the resulting misadjustments are quite different. NLMS with $\bar{\mu} = 1$ has a misadjustment $\mathcal{M}$ of 1.1164, whereas APA with $M = 2$ has a misadjustment $\mathcal{M}$ of 0.2904. In other words, the steady-state error of APA with $M = 2$ is at least 2 dB less than the steady-state error of NLMS with $\bar{\mu} = 1$, whereas their convergence rates are comparable. APA with $M = 1$ (not shown to avoid clutter) has a misadjustment $\mathcal{M}$ of 0.2269 and converges almost as fast as NLMS with $\bar{\mu} = 1$. We note that the (experimental) misadjustment has some dependence on M (misadjustment increases as M increases). This increase in misadjustment with M has been reported in earlier papers [11]–[13]. However, the misadjustment has a stronger dependence on step size than on M. This suggests that it would be better to use APA to get improved convergence than to use NLMS with large step size.

Fig. 11 depicts the dependence of experimental misadjustment on M. Here, the misadjustments for different values of M and different step-size constants $\bar{\mu}$ are shown. We see that the dependence on M increases as the step size is increased. For small values of step size, the misadjustment does not change much with M. This supports our hypothesis that the misadjustment, as shown in (32), is independent of M since we neglected the dependence of $\tilde{\mathbf{w}}_n$ on past measurement noise while going from (11) to (12). As the step size is decreased, the dependence of $\tilde{\mathbf{w}}_n$ on past measurement noise decreases, and hence, neglecting this dependence does not introduce "too much" error. Thus, our Observation 5 that the misadjustment for APA does not depend on M holds as long as the data and parameters satisfy our assumptions.

V. Conclusion

The APA class of algorithms provides an improvement in convergence rate over NLMS, especially for colored input signals. We analyzed the convergence behavior of APA based on the simplifying assumptions that the input vectors are independent and have a discrete angular orientation. A theoretical expression for the convergence behavior of the mean-squared error is derived. As the signal color, input vector delay, and/or step sizes tend toward satisfying the independence assumption, the simulated results tend to the theoretical results, whereas there is a mismatch otherwise. The convergence rate is exponential, and it improves with an increase in the number of input signal vectors used for adaptation. However, the *rate* of improvement in time-to-steady-state diminishes as the number of input vectors used for adaptation increases.

For white input, the mean squared error drops by 20 dB in about $5N/(M+1)$ iterations, where N is the number of taps in the adaptive filter, and M is the number of vectors used for adaptation. Although we show that in theory, the misadjustment of the APA class is independent of the number of vectors used for adaptation, simulation results show a weak dependence. Thus, APA provides a way to increase the convergence rate without compromising too much on misadjustment. Simulation results corroborate our findings.

Appendix

When $\bar{\mu} = 1$, the weight update generated by APA is the vector that is as close as possible to the current weight vector while setting the most recent $(M + 1)$ *a posteriori* error estimates to zero [2]. That is

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \Delta\mathbf{w}_n \tag{49}$$

where $\Delta\mathbf{w}_n$ is the minimum-norm solution to

$$\mathbf{X}_n^T \Delta\mathbf{w}_n = \mathbf{e}_n. \tag{50}$$

In the above equation, $\mathbf{X}_n = [\mathbf{x}_n\ \mathbf{x}_{n-1}\ \cdots\ \mathbf{x}_{n-M}], \mathbf{e}_n = [e_n\ \tilde{e}_n^1\ \cdots\ \tilde{e}_n^M]^T, e_n = d_n - \mathbf{x}_n^T\mathbf{w}_n$, and $\tilde{e}_n^k = d_{n-k} -$

$\mathbf{x}_{n-k}^T\mathbf{w}_n$. Since $\Delta\mathbf{w}_n$ is the minimum-norm solution of (50), it is the unique solution of (50) that lies in the space spanned by the columns of $\mathbf{X}_n$. APA usually solves for $\Delta\mathbf{w}_n$ using the matrix equation

$$\Delta\mathbf{w}_n = \mathbf{X}_n[\mathbf{X}_n^T\mathbf{X}_n]^{-1}\mathbf{e}_n. \tag{51}$$

Observe that the above solution lies in the space spanned by the columns of $\mathbf{X}_n$. Simple algebra shows that $\mathbf{w}_{n+1}$ obtained using (49) and (51) sets the most recent $(M+1)$ *a posteriori* error estimates to zero. That is

$$\mathbf{X}_n^T\mathbf{w}_{n+1} = [d_n \;\; d_{n-1} \;\; \cdots \;\; d_{n-M}]^T. \tag{52}$$

NLMS-OCF, on the other hand, finds the weight update by setting "one *a posteriori* estimation error at a time to zero," as explained below. NLMS-OCF begins by setting the *a posteriori* estimation error at n to zero while keeping the norm of the increment in weights to a minimum. That is, it finds the weight $\mathbf{w}_n^1$ such that $\|\mathbf{w}_n^1 - \mathbf{w}_n\|$ is minimized subject to $d_n - \mathbf{x}_n^T\mathbf{w}_n^1 = 0$. This solution is given by

$$\mathbf{w}_n^1 = \mathbf{w}_n + \mu_0\mathbf{x}_n \tag{53}$$

where $\mu_0 = (e_n/\mathbf{x}_n^T\mathbf{x}_n)$, and $e_n = d_n - \mathbf{x}_n^T\mathbf{w}_n$.

Next, NLMS-OCF finds the weight $\mathbf{w}_n^2$ that forces the *a posteriori* estimation error at $(n-1)$ to zero while maintaining the zero *a posteriori* estimation error at n and keeping the norm of the increment in weights to a minimum. That is, find the weight $\mathbf{w}_n^2$ such that $\|\mathbf{w}_n^2 - \mathbf{w}_n\|$ is minimized subject to $d_n - \mathbf{x}_n^T\mathbf{w}_n^2 = 0$, and $d_{n-1} - \mathbf{x}_{n-1}^T\mathbf{w}_n^2 = 0$. If the increment in weights $(\mathbf{w}_n^2 - \mathbf{w}_n^1)$ is orthogonal to $\mathbf{x}_n$, then $d_n - \mathbf{x}_n^T\mathbf{w}_n^2 = d_n - \mathbf{x}_n^T\mathbf{w}_n^1 = 0$. Thus, the first constraint is satisfied if the weight increment is orthogonal to $\mathbf{x}_n$. Hence, we decompose $\mathbf{x}_{n-1}$ into a component along $\mathbf{x}_n$ and a component $\mathbf{x}_n^1$ that is orthogonal to $\mathbf{x}_n$. We increment the weights along $\mathbf{x}_n^1$ such that the second constraint is satisfied. This solution is given by

$$\begin{aligned}\mathbf{w}_n^2 &= \mathbf{w}_n^1 + \mu_1\mathbf{x}_n^1 \\ &= \mathbf{w}_n + \mu_0\mathbf{x}_n + \mu_1\mathbf{x}_n^1\end{aligned} \tag{54}$$

where $\mu_0 = (e_n^1/{\mathbf{x}_n^1}^T\mathbf{x}_n^1)$ and $e_n^1 = d_{n-1} - \mathbf{x}_{n-1}^T\mathbf{w}_n^1$.

The above process is repeated until each of the most recent $(M+1)$ *a posteriori* errors is forced to zero. We describe here the general step that forces the *a posteriori* estimation error at $(n-k)$ to zero, where $k \in \{1,2,\ldots,M\}$. Here, we find the weight $\mathbf{w}_n^k$ such that $\|\mathbf{w}_n^k - \mathbf{w}_n\|$ is minimized subject to $d_{n-r} - \mathbf{x}_{n-r}^T\mathbf{w}_n^k = 0$ for $r = 0,1,\ldots k-1$, and $d_{n-k} - \mathbf{x}_{n-k}^T\mathbf{w}_n^k = 0$. If the increment in weights $(\mathbf{w}_n^k - \mathbf{w}_n^{k-1})$ is orthogonal to $\mathbf{x}_n, \mathbf{x}_{n-1}, \ldots, \mathbf{x}_{n-(k-1)}$, then $d_{n-r} - \mathbf{x}_{n-r}^T\mathbf{w}_n^k = d_{n-r} - \mathbf{x}_{n-r}^T\mathbf{w}_n^{k-1} = 0$ for $r = 0,1,\ldots k-1$. Thus, the first k constraints are satisfied if the increment is orthogonal to $\mathbf{x}_n, \mathbf{x}_{n-1}, \ldots, \mathbf{x}_{n-(k-1)}$. Hence, we decompose $\mathbf{x}_{n-k}$ into a component that is in the span of $\mathbf{x}_n, \mathbf{x}_{n-1}, \ldots, \mathbf{x}_{n-(k-1)}$ and a component $\mathbf{x}_n^k$ that is orthogonal to $\mathbf{x}_n, \mathbf{x}_{n-1}, \ldots, \mathbf{x}_{n-(k-1)}$. We increment the weights along $\mathbf{x}_n^k$ such that the last constraint is satisfied. This solution is given by

$$\begin{aligned}\mathbf{w}_n^{k+1} &= \mathbf{w}_n^k + \mu_1\mathbf{x}_n^k \\ &= \mathbf{w}_n + \mu_0\mathbf{x}_n + \mu_1\mathbf{x}_n^1 + \cdots + \mu_k\mathbf{x}_n^k\end{aligned} \tag{55}$$

where $\mu_k = (e_n^k/{\mathbf{x}_n^k}^T\mathbf{x}_n^k)$, and $e_n^k = d_{n-k} - \mathbf{x}_{n-k}^T\mathbf{w}_n^k$.

Thus, the weight update that forces the most recent $(M+1)$ *a posteriori* estimation errors to zero is given by

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \mu_0\mathbf{x}_n + \mu_1\mathbf{x}_n^1 + \cdots + \mu_M\mathbf{x}_n^M \tag{56}$$

where $M+1$ is the number of input vectors used for adaptation, $\mathbf{x}_n$ is the input vector at the nth instant, $\mathbf{x}_n^k$, for $k = 1,2,\ldots,M$, is the component of $\mathbf{x}_{n-k}$ that is orthogonal to $\mathbf{x}_n, \mathbf{x}_{n-1}, \mathbf{x}_{n-2}, \ldots, \mathbf{x}_{n-(k-1)}$, and μ_k, for $k = 0,1,\ldots,M$ is chosen as in

$$\mu_k = \begin{cases} \dfrac{e_n}{\mathbf{x}_n^T\mathbf{x}_n}, & \text{for } k=0, \quad \text{if } \|\mathbf{x}_n\| \neq 0 \\ \dfrac{e_n^k}{{\mathbf{x}_n^k}^T\mathbf{x}_n^k} & \text{for } k=1,2,\ldots,M, \quad \text{if } \|\mathbf{x}_n^k\| \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{57}$$

where

$$\begin{aligned} e_n &= d_n - \mathbf{x}_n^T\mathbf{w}_n \\ e_n^k &= d_{n-k} - \mathbf{x}_{n-k}^T\mathbf{w}_n^k, \quad \text{for } k=1,2,\ldots,M, \quad \text{and} \\ \mathbf{w}_n^k &= \mathbf{w}_n + \mu_0\mathbf{x}_n + \mu_1\mathbf{x}_n^1 + \cdots + \mu_{k-1}\mathbf{x}_n^{k-1}. \end{aligned} \tag{58}$$

Observe from (56) that the increment in weight lies in the space spanned by the columns of $\mathbf{X}_n$. Furthermore, the updated weight satisfies (52). Equivalently, the weight increment satisfies (50). Since the minimum-norm solution to (50) is the unique solution of (50) that is in the space spanned by the columns of $\mathbf{X}_n$, the weight updates generated by APA and by NLMS-OCF with $D = 1$ are identical.

As is usually done in APA, the above algorithm can be generalized by introducing a constant $\bar{\mu}$, which is usually referred to as the step size. This generalization, along with the modifications needed for the complex case, results in the update equations (1)–(3).

REFERENCES

[1] S. Haykin, *Adaptive Filter Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1991.

[2] D. R. Morgan and S. G. Kratzer, "On a class of computationally efficient, rapidly converging, generalized NLMS algorithms," *IEEE Signal Processing Lett.*, vol. 3, pp. 245–247, Aug. 1996.

[3] S. G. Sankaran and A. A. (Louis) Beex, "Normalized LMS algorithm with orthogonal correction factors," in *Proc. Thirty-First Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, Nov. 2–5, 1997, pp. 1670–1673.

[4] D. T. M. Slock, "On the convergence behavior of the LMS and the normalized LMS algorithms," *IEEE Trans. Signal Processing*, vol. 41, pp. 2811–2825, Sept. 1993.

[5] M. Tarrab and A. Feuer, "Convergence and performance analysis of the normalized LMS algorithm with uncorrelated Gaussian data," *IEEE Trans. Inform. Theory*, vol. 34, pp. 680–691, July 1988.

[6] Z. Pritzker and A. Feuer, "Variable length stochastic gradient algorithm," *IEEE Trans. Signal Processing*, vol. 39, pp. 997–1001, Apr. 1991.

[7] K. Ozeki and T. Umeda, "An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties," *Electron. Commun. Jpn.*, vol. 67-A, no. 5, pp. 19–27, 1984.

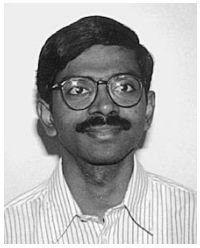
[8] S. L. Gay and S. Tavathia, "The fast affine projection algorithm," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Detroit, MI, May 8–12, 1995, pp. 3023–3026.

[9] S. Shimauchi and S. Makino, "Stereo projection echo canceler with true echo path estimation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Detroit, MI, May 8–12, 1995, pp. 3059–3062.

[10] Y. Kaneda, M. Tanaka, and J. Kojima, "An adaptive algorithm with fast convergence for multi-input sound control," in *Proc. Active*, Newport Beach, CA, July 6–8, 1995, pp. 993–1004.

[11] M. Rupp, "A family of adaptive filter algorithms with decorrelating properties," *IEEE Trans. Signal Processing*, vol. 46, pp. 771–775, Mar. 1998.

[12] D. Slock, "The block underdetermined covariance (BUC) fast transversal filter (FTF) algorithm for adaptive filtering," in *Proc. Twenty-Sixth Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, 1992, pp. 550–554.
[13] B. Baykal, O. Tanrikulu, and A. G. Constantinides, "Asymptotic analysis of the underdetermined recursive least-squares algorithm," in *Proc. EUSIPCO*, Trieste, Italy, 1996, pp. 1397–1400.
[14] S. G. Sankaran and A. A. Beex, "Fast generalized affine projection algorithm," *Proc. Int. J. Adaptive Contr. Signal Process.*, Feb. 2000.
[15] ——, "Stereophonic acoustic echo cancellation using NLMS with orthogonal correction factors," in *Proc. Int. Workshop Acoust. Echo Noise Contr.*, Pocono Manor, PA, Sept. 1999, pp. 40–43.



Sundar G. Sankaran (S'96) received the B.Eng. degree in electronics and communication engineering in 1992 from Anna University, Madras, India, and the M.Sc. and Ph.D. degrees in electrical engineering in 1996 and 1999, respectively, from Virginia Tech, Blacksburg.

From 1992 to 1994, he was at Infosys Technologies Limited, Bangalore, India, as a Systems Analyst, where he worked on digital signal processing hardware design and embedded software development. Since 1995, he has been a Graduate Research Assistant with the DSP Research Laboratory at Virginia Tech. His research interests are primarily in the area of digital signal processing and its applications.

A. A. (Louis) Beex (SM'86) received the Ingenieur degree from Technical University Eindhoven, the Netherlands, in 1974 and the Ph.D. degree from Colorado State University, Fort Collins, in 1979, both in electrical engineering.

From 1976 to 1978, he was a Staff Research Engineer at Starkey Laboratories, Minneapolis, MN, applying DSP to hearing instrumentation. He has been a member of the faculty of the Department of Electrical and Computer Engineering at Virginia Tech, Blacksburg, for the past two decades, is Director of the DSP Research Laboratory at Virginia Tech, and runs DSP Consultants, a small enterprise. His interests lie in the design, analysis, and implementation aspects of DSP algorithms for various applications.

Dr. Beex is a past Associate Editor of the IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING.