## 3. Convergence Behavior of the APA Class Of Algorithms

In this section we analyze the convergence behavior of the NLMS-OCF algorithm and the algorithms equivalent to NLMS-OCF such as the Affine Projection Algorithm (APA), the Partial Rank Algorithm (PRA) and the Generalized Optimal Block Algorithm (GOBA). In the sequel, we will refer to this entire class of algorithms as affine projection algorithms, since APA is the earliest among these algorithms and since the name APA is more widely used in the existing literature than the other names. We use the weight update equation of the NLMS-OCF algorithm for our analysis, since it is more general than in the other algorithms and since the NLMS-OCF update equation is conducive to the analysis that follows. 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.

While 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 [20, 40, 41]. In this chapter, 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. While 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.

### 3.1 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

$$\mathbf{R} = E\left[\mathbf{x}_n \mathbf{x}_n^H\right] = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^H \tag{3.1}$$

where $\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, ..., \lambda_N)$ and $\mathbf{V} = (\boldsymbol{\nu}_1 \quad \boldsymbol{\nu}_2 \quad ... \quad \boldsymbol{\nu}_N)$. Here, $\lambda_1, \lambda_2, ..., \lambda_N$ are the eigenvalues of $\mathbf{R}$ and $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, ..., \boldsymbol{\nu}_N$ 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

$$\begin{aligned} e_n &= d_n - w^{0H}\mathbf{x}_n \\ &\equiv \varepsilon_n \end{aligned} \tag{3.2}$$

inherits the properties of the measurement noise $\varepsilon_n$, which is a zero mean white noise of variance $\xi^0$ 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,

$$\mathbf{x}_n = s_n \, r_n \, \mathbf{v}_n \tag{3.3a}$$

where

$$\begin{cases} P\{s_n = \pm 1\} = \dfrac{1}{2} \\ r_n \sim \|\mathbf{x}_n\| \\ P\{\mathbf{v}_n = \boldsymbol{\nu}_i\} = p_i = \dfrac{\lambda_i}{tr(\mathbf{R})}, \quad i = 1,2,...,N. \end{cases} \tag{3.3b}$$

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), 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 $k = 1,2,...,M$, becomes redundant. Hence, (2.6), (2.7), and (2.9) can be rewritten as shown in (3.4), (3.5), and (3.6).

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \mu_0 \mathbf{x}_n + \mu_1 \mathbf{x}_{n-D} + ... + \mu_M \mathbf{x}_{n-MD} \tag{3.4}$$

$$\mu_k = \begin{cases} \dfrac{\bar{\mu} e_n^*}{\mathbf{x}_n^H \mathbf{x}_n} & for \;\; k = 0, \; if \;\; \|\mathbf{x}_n\| \neq 0 \\ \dfrac{\bar{\mu} e_n^{k*}}{\mathbf{x}_{n-kD}^H \mathbf{x}_{n-kD}} & for \;\; k = 1,2,...,M, \; if \;\; \mathbf{x}_{n-kD} \perp \mathbf{x}_{n-iD} \;\; \forall i < k \\ 0 & otherwise \end{cases} \tag{3.5}$$

$$\begin{aligned} e_n &= d_n - \mathbf{w}_n^H \mathbf{x}_n \;, \; and \\ e_n^k &= d_{n-kD} - \mathbf{w}_n^H \mathbf{x}_{n-kD}, \;\; for \; k = 1,2,...,M. \end{aligned} \tag{3.6}$$

(Using (A3), $\mathbf{w}_n^{kH} \mathbf{x}_{n-kD} = \left(\mathbf{w}_n + \mu_0 \mathbf{x}_n + \mu_1 \mathbf{x}_{n-D} + ... + \mu_{k-1} \mathbf{x}_{n-(k-1)D}\right)^H \mathbf{x}_{n-kD} = \mathbf{w}_n^H \mathbf{x}_{n-kD}$, since $\mathbf{x}_{n-kD} \perp \mathbf{x}_{n-iD} \;\; \forall i < k$. Hence, (2.9b) can be modified to the form shown in (3.6).)

To analyze the convergence behavior of (3.4), firstly, 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 (3.2), 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 (3.4) and (3.5), the adaptation equation in error form can be obtained as:

$$\tilde{\mathbf{w}}_{n+1} = \left[ \mathbf{I} - \sum_{j \in J_n} \bar{\mu} \frac{\mathbf{x}_{n-jD} \mathbf{x}_{n-jD}^H}{\mathbf{x}_{n-jD}^H \mathbf{x}_{n-jD}} \right] \tilde{\mathbf{w}}_n - \sum_{l \in J_n} \bar{\mu} \frac{\varepsilon_{n-lD}^* \mathbf{x}_{n-lD}}{\mathbf{x}_{n-lD}^H \mathbf{x}_{n-lD}}, \tag{3.7}$$

where $J_n \subseteq \{0,1,2,...,M\}$ 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 (3.7) is in a form suitable for convergence analysis. In the absence of noise $\varepsilon_n$, (3.7) 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 $\mathrm{cov}(\tilde{\mathbf{w}}_n)$ 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 (3.7), the covariance of the weight error vector $\tilde{\mathbf{w}}_n$ is given by:

$$\begin{aligned}\mathrm{cov}(\tilde{\mathbf{w}}_{n+1}) = {} & E\left(\left[\mathbf{I}-\sum_{j\in J_n}\bar{\mu}\frac{\mathbf{x}_{n-jD}\mathbf{x}_{n-jD}^H}{\mathbf{x}_{n-jD}^H\mathbf{x}_{n-jD}}\right]\tilde{\mathbf{w}}_n\tilde{\mathbf{w}}_n^H\left[\mathbf{I}-\sum_{l\in J_n}\bar{\mu}\frac{\mathbf{x}_{n-lD}\mathbf{x}_{n-lD}^H}{\mathbf{x}_{n-lD}^H\mathbf{x}_{n-lD}}\right]\right)\\ & + E\left(\left[\sum_{j\in J_n}\bar{\mu}\frac{\varepsilon_{n-jD}^*\mathbf{x}_{n-jD}}{\mathbf{x}_{n-jD}^H\mathbf{x}_{n-jD}}\right]\left[\sum_{l\in J_n}\bar{\mu}\frac{\varepsilon_{n-lD}\mathbf{x}_{n-lD}^H}{\mathbf{x}_{n-lD}^H\mathbf{x}_{n-lD}}\right]\right)\\ & - E\left(\left[\mathbf{I}-\sum_{j\in J_n}\bar{\mu}\frac{\mathbf{x}_{n-jD}\mathbf{x}_{n-jD}^H}{\mathbf{x}_{n-jD}^H\mathbf{x}_{n-jD}}\right]\tilde{\mathbf{w}}_n\left[\sum_{l\in J_n}\bar{\mu}\frac{\varepsilon_{n-lD}\mathbf{x}_{n-lD}^H}{\mathbf{x}_{n-lD}^H\mathbf{x}_{n-lD}}\right]\right)\\ & - E\left(\left[\sum_{j\in J_n}\bar{\mu}\frac{\varepsilon_{n-jD}^*\mathbf{x}_{n-jD}}{\mathbf{x}_{n-jD}^H\mathbf{x}_{n-jD}}\right]\tilde{\mathbf{w}}_n^H\left[\mathbf{I}-\sum_{l\in J_n}\bar{\mu}\frac{\mathbf{x}_{n-lD}\mathbf{x}_{n-lD}^H}{\mathbf{x}_{n-lD}^H\mathbf{x}_{n-lD}}\right]\right)\end{aligned} \tag{3.8}$$

If the dependency of $\tilde{\mathbf{w}}_n$ on past measurement noise is neglected, using that $\varepsilon_n$ 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 (3.8) as

$$\begin{aligned}\mathrm{cov}(\tilde{\mathbf{w}}_{n+1}) = {} & E\left(\left[\mathbf{I}-\sum_{j\in J_n}\bar{\mu}\frac{\mathbf{x}_{n-jD}\mathbf{x}_{n-jD}^H}{\mathbf{x}_{n-jD}^H\mathbf{x}_{n-jD}}\right]\mathrm{cov}(\tilde{\mathbf{w}}_n)\left[\mathbf{I}-\sum_{l\in J_n}\bar{\mu}\frac{\mathbf{x}_{n-lD}\mathbf{x}_{n-lD}^H}{\mathbf{x}_{n-lD}^H\mathbf{x}_{n-lD}}\right]\right)\\ & + E\left(\bar{\mu}^2\sum_{j\in J_n}\left|\varepsilon_{n-jD}\right|^2\frac{\mathbf{x}_{n-jD}\mathbf{x}_{n-jD}^H}{\left\|\mathbf{x}_{n-jD}\right\|^2\mathbf{x}_{n-jD}^H\mathbf{x}_{n-jD}}\right)\end{aligned} \tag{3.9}$$

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

$$\begin{aligned}\frac{\mathbf{x}_{n-jD}\mathbf{x}_{n-jD}^H}{\mathbf{x}_{n-jD}^H\mathbf{x}_{n-jD}} &= \frac{s_{n-jD}r_{n-jD}\mathbf{v}_{n-jD}\mathbf{v}_{n-jD}^H r_{n-jD}s_{n-jD}}{s_{n-jD}^2 r_{n-jD}^2\left\|\mathbf{v}_{n-jD}\right\|^2} \\ &= \mathbf{v}_{n-jD}\mathbf{v}_{n-jD}^H\end{aligned} \tag{3.10}$$

where $\mathbf{v}_{n-jD}$ is one of $\{\mathbf{v}_1,\mathbf{v}_2,...,\mathbf{v}_N\}$. Note that the above result is independent of the norm of $\mathbf{x}_{n-jD}$. Now, substituting (3.10) into (3.9) we get,

$$\begin{aligned}\mathrm{cov}(\tilde{\mathbf{w}}_{n+1}) = E\left(\left[\mathbf{I}-\sum_{j\in J_n}\bar{\mu}\mathbf{v}_{n-jD}\mathbf{v}_{n-jD}^H\right]\mathrm{cov}(\tilde{\mathbf{w}}_n)\left[\mathbf{I}-\sum_{l\in J_n}\bar{\mu}\mathbf{v}_{l-jD}\mathbf{v}_{l-jD}^H\right]\right) \\ + E\left(\bar{\mu}^2\sum_{j\in J_n}\left|\varepsilon_{n-jD}\right|^2\frac{1}{r_{n-jD}^2}\mathbf{v}_{n-jD}\mathbf{v}_{n-jD}^H\right)\end{aligned} \tag{3.11}$$

Since $\varepsilon_n$ is independent of $\mathbf{x}_n$ and *r* is independent of **v**, from (A2) and (A3) respectively, we can rewrite (3.11) as

$$\mathrm{cov}(\tilde{\mathbf{w}}_{n+1}) = E\left(\left[\mathbf{I}-\sum_{k\in K_n}\bar{\mu}\mathbf{v}_k\mathbf{v}_k^H\right]\mathrm{cov}(\tilde{\mathbf{w}}_n)\left[\mathbf{I}-\sum_{l\in K_n}\bar{\mu}\mathbf{v}_k\mathbf{v}_k^H\right]\right)+\bar{\mu}^2\xi_0 E\left(\frac{1}{r^2}\right)E\left(\sum_{k\in K_n}\mathbf{v}_k\mathbf{v}_k^H\right) \tag{3.12}$$

where

$$K_n = \left\{k : \exists j \in J_n \ni \frac{\mathbf{x}_{n-jD}\mathbf{x}_{n-jD}^H}{\mathbf{x}_{n-jD}^H\mathbf{x}_{n-jD}} = \mathbf{v}_k\mathbf{v}_k^H\right\} \subseteq \{1,2,...,N\}. \tag{3.13}$$

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

$$\left[\mathbf{V}^H\mathrm{cov}(\tilde{\mathbf{w}}_n)\mathbf{V}\right]_{ii} = \mathbf{v}_i^H\mathrm{cov}(\tilde{\mathbf{w}}_n)\mathbf{v}_i = \tilde{\lambda}_{n,i}. \tag{3.14}$$

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

---

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

With the above notation, the pre- and post-multiplication of (3.12) by $\mathbf{v}_i^H$ and $\mathbf{v}_i$ respectively results in

$$\begin{aligned}\tilde{\lambda}_{n+1,i} = E\left(\mathbf{v}_i^H\left[\mathbf{I}-\sum_{k\in K_n}\bar{\mu}\mathbf{v}_k\mathbf{v}_k^H\right]\mathrm{cov}(\tilde{\mathbf{w}}_n)\left[\mathbf{I}-\sum_{k\in K_n}\bar{\mu}\mathbf{v}_k\mathbf{v}_k^H\right]\mathbf{v}_i\right)\\ +\bar{\mu}^2\xi^0 E\left(\frac{1}{r^2}\right)E\left(\mathbf{v}_i^H\left[\sum_{k\in K_n}\mathbf{v}_k\mathbf{v}_k^H\right]\mathbf{v}_i\right)\end{aligned} \tag{3.15}$$

From the orthonormality of the $\mathbf{v}_k$'s,

$$\mathbf{v}_i^H\sum_{k\in K_n}\mathbf{v}_k\mathbf{v}_k^H = \begin{cases}\mathbf{v}_i^H, & if\ i\in K_n\\ \mathbf{0}, & if\ i\notin K_n.\end{cases} \tag{3.16}$$

Using the above result, (3.15) can be rewritten as

$$\begin{aligned}\tilde{\lambda}_{n+1,i} &= \mathbf{v}_i^H\,\mathrm{cov}(\tilde{\mathbf{w}}_n)\mathbf{v}_i + E\left(\mathbf{v}_i^H\left[\sum_{k\in K_n}\bar{\mu}\mathbf{v}_k\mathbf{v}_k^H\right]\mathrm{cov}(\tilde{\mathbf{w}}_n)\left[\sum_{l\in K_n}\bar{\mu}\mathbf{v}_l\mathbf{v}_l^H\right]\mathbf{v}_i\right)+\\ &\quad -E\left(\mathbf{v}_i^H\,\mathrm{cov}(\tilde{\mathbf{w}}_n)\left[\sum_{l\in K_n}\bar{\mu}\mathbf{v}_l\mathbf{v}_l^H\right]\mathbf{v}_i\right)-E\left(\mathbf{v}_i^H\left[\sum_{k\in K_n}\bar{\mu}\mathbf{v}_k\mathbf{v}_k^H\right]\mathrm{cov}(\tilde{\mathbf{w}}_n)\,\mathbf{v}_i\right)\\ &\quad +\bar{\mu}^2\,\xi^0 E\left(\frac{1}{r^2}\right)E\left(\mathbf{v}_i^H\left[\sum_{k\in K_n}\mathbf{v}_k\mathbf{v}_k^H\right]\mathbf{v}_i\right)\\ &= \tilde{\lambda}_{n,i}\left[1-\bar{\mu}(2-\bar{\mu})P(i\in K_n)\right]+\bar{\mu}^2\xi^0 E\left(\frac{1}{r^2}\right)P(i\in K_n)\end{aligned} \tag{3.17}$$

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,…,*N*, where the probability of drawing the ball marked *j* is $p_j$. Hence

$$P(i\in K_n) = 1-(1-p_i)^{M+1}. \tag{3.18}$$

By substituting (3.18) into (3.17), we get

$$\tilde{\lambda}_{n+1,i} = (1-\alpha\beta_i)\tilde{\lambda}_{n,i}+\bar{\mu}^2\xi^0 E\left(\frac{1}{r^2}\right)\beta_i \tag{3.19}$$

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

The following observations can be made from (3.19):

*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

$$e_n = \tilde{\mathbf{w}}_n^H \mathbf{x}_n + \varepsilon_n. \tag{3.20}$$

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

$$\begin{aligned} \xi_n &= \xi^0 + E\left( \left\| \tilde{\mathbf{w}}_n^H \mathbf{x}_n \right\|^2 \right) \\ &= \xi^0 + tr\left[\mathbf{R}\,\mathrm{cov}(\tilde{\mathbf{w}}_n)\right] \\ &= \xi^0 + tr\left[\mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^H\,\mathrm{cov}(\tilde{\mathbf{w}}_n)\right] \\ &= \xi^0 + tr\left[\boldsymbol{\Sigma}\ \mathbf{V}^H \mathrm{cov}(\tilde{\mathbf{w}}_n)\mathbf{V}\right] \\ &= \xi^0 + \sum_{i=1}^{N} \lambda_i\, \tilde{\lambda}_{n,i} \end{aligned} \tag{3.21}$$

From (3.21), 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 $0 \le (1-\alpha\beta_i) < 1$; this guarantees the convergence of $\tilde{\lambda}_{n,i}$ in (22). If $\bar{\mu} \notin (0,2)$, then $\alpha \le 0$ and $(1-\alpha\beta_i) \ge 1$; 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 (3.21) is given by

$$\lim_{n\to\infty} \tilde{\lambda}_{n,i} = \frac{\bar{\mu}}{2-\bar{\mu}} \xi^0 E\left(\frac{1}{r^2}\right). \tag{3.22}$$

Combining (3.21) and (3.22), the steady-state (final) mean-squared error is given by

$$\xi_\infty = \xi^0 \left[ 1 + \frac{\bar{\mu}}{2-\bar{\mu}} E\left(\frac{1}{r^2}\right) tr(\mathbf{R}) \right] < \infty \tag{3.23}$$

Using (3.21), 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 (3.7) results in:

$$E(\tilde{\mathbf{w}}_{n+1}) = E(\tilde{\mathbf{w}}_n) - E\left( \bar{\mu} \sum_{k \in K_n} \boldsymbol{\nu}_k \boldsymbol{\nu}_k^H \tilde{\mathbf{w}}_n \right). \tag{3.24}$$

Here we used (3.13) 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 $\varepsilon_n$ vanishes.

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

$$\boldsymbol{\rho}_n \equiv \mathbf{V}^H E(\tilde{\mathbf{w}}_n). \tag{3.25}$$

Therefore,

$$\rho_{n,i} = \boldsymbol{\nu}_i^H E(\tilde{\mathbf{w}}_n) = E(\boldsymbol{\nu}_i^H \tilde{\mathbf{w}}_n). \tag{3.26}$$

Using this notation, pre-multiplication of (3.24) by $\boldsymbol{\nu}_i^H$ results in

$$\rho_{n+1,i} = \rho_{n,i} - E\left( \bar{\mu} \boldsymbol{\nu}_i^H \sum_{k \in K_n} \boldsymbol{\nu}_k \boldsymbol{\nu}_k^H \tilde{\mathbf{w}}_n \right). \tag{3.27}$$

Using (3.16) and (3.18), (3.27) can be rewritten as

$$\rho_{n+1,i} = (1 - \bar{\mu}\beta_i)\rho_{n,i}. \tag{3.28}$$

From (3.28) 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 $\{\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, ..., \boldsymbol{\nu}_N\}$ 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.

*Observation 2:* *The misadjustment of the APA class is independent of M.* Using (3.23), the misadjustment, defined as the ratio of excess mean-squared error to minimum mean-squared error, equals

$$\mathcal{M} = \frac{\xi_\infty - \xi^0}{\xi^0} = \frac{\bar{\mu}}{2-\bar{\mu}} E\left(\frac{1}{r^2}\right) tr(\mathbf{R}) \tag{3.29}$$

Note the independence of (3.29) 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 (3.29) of *M* is, perhaps, due to the fact that we neglected dependence of $\tilde{\mathbf{w}}_n$ on past measurement noise while going from (3.8) to (3.9). Simulation results indicate a "weak" dependence of misadjustment on *M*. Later, we will show that 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 3:* *The convergence behavior of the mean-squared error $\xi_n$ for the noiseless case, viz. $\xi^0 = 0$, is exponential,* as given in (3.35). 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 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 **R**. For example,

$$\mathbf{w}^0 = \sqrt{\frac{\sigma_d^2 - \xi^0}{tr(\mathbf{R})}}\mathbf{V}1_N, \tag{3.30}$$

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

$$\text{cov}(\tilde{w}_0) = E(\tilde{\mathbf{w}}_0 \tilde{\mathbf{w}}_0^H) = \mathbf{w}^0 \mathbf{w}^{0H} = \frac{\sigma_d^2}{tr(\mathbf{R})}\mathbf{V}1_N 1_N^H \mathbf{V}^H. \tag{3.31}$$

Using the fact that **V** is unitary, it follows that

$$\mathbf{V}^H \mathrm{cov}(\tilde{w}_0)\mathbf{V} = \frac{\sigma_d^2}{tr(\mathbf{R})} 1_N 1_N^H . \tag{3.32}$$

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

$$\tilde{\lambda}_{0,i} = \sigma_d^2 / tr(\mathbf{R}) \; \forall i . \tag{3.33}$$

Solving (3.19), using (3.33) as the initial condition, and substituting the solution in (3.21), we get the mean-squared error as

$$\xi_n = \sum_{i=1}^{N} \lambda_i (1 - \alpha\beta_i)^n \frac{\sigma_d^2}{tr(\mathbf{R})} . \tag{3.34}$$

From (A3), $\lambda_i / tr(\mathbf{R}) = p_i$, so that we can rewrite (3.34) as follows

$$\xi_n = \sigma_d^2 \sum_{i=1}^{N} (1 - \alpha\beta_i)^n \, p_i . \tag{3.35}$$

Hence, (3.35) describes the theoretical convergence behavior of the APA class of algorithms under noise-free conditions.

<u>*Observation 4:*</u> *APA converges faster than NLMS; as more input vectors are used, the convergence rate itself improves whereas the rate of this improvement decreases.* From (3.19), 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 $\alpha$, and hence the convergence rates, are the same for step sizes $\bar{\mu}$ and $2 - \bar{\mu}$ for $\bar{\mu} \in (0,2)$. However, from (3.19), the misadjustment 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 (3.19), faster convergence occurs for values of $(1 - \alpha\beta_i)$ closer to 0 (equivalently $\alpha$ and $\beta_i$ 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 $\beta_i$ gets closer to 1. This explains the faster convergence of APA over NLMS. Figure 3.1 shows a plot of the convergence rate factor $(1 - \alpha\beta_i)$ for different values of *M* and different values of $p_i$, with $\bar{\mu} = 1$.

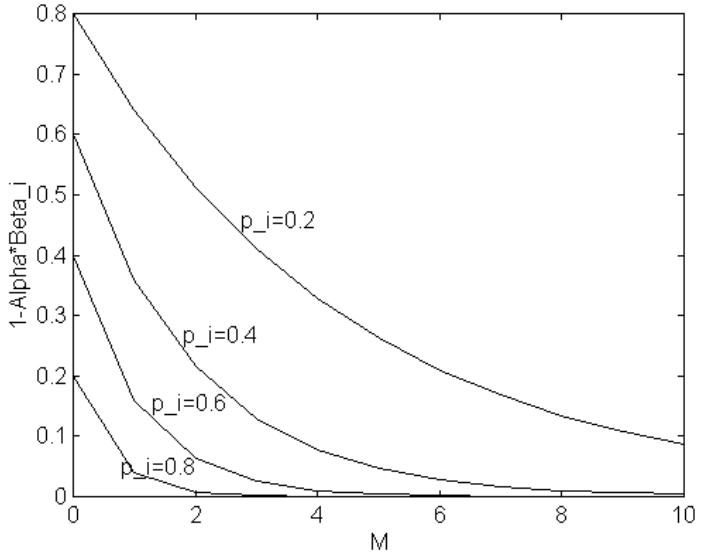


**Figure 3.1 Dependence of the convergence rate factor $(1-\alpha\beta_i)$ on *M*.**

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 $\eta_i^M$ for some $\eta_i < 1$. Hence, for large enough values of *n*, with $\overline{p}_i$ denoting the total probability mass associated with the largest of the $\eta_i$, (3.35) can be approximated as

$$\xi_n \approx \sigma_d^2 (\max_i \eta_i)^{Mn} \overline{p}_i \tag{3.36}$$

Equivalently,

$$\xi_{n,dB} = 10\log_{10}\sigma_d^2 \overline{p}_i + 10Mn\log_{10}(\max_i \eta_i) \tag{3.37}$$

Thus, for large enough *n*, the slope of the learning curve (plot showing mean-squared error in dB 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 [19].

<u>*Observation 5:*</u> *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 and that $\overline{\mu} = 1$. Since the input is white, all the $p_i$'s are equal. That is,

$$p_i = \frac{1}{N} \quad for\ i = 1,2,...,N. \tag{3.38}$$

Therefore, the convergence rate factor for white noise, corresponding to $\bar{\mu} = 1$, is

$$(1 - \alpha\beta_i) = \left(1 - \frac{1}{N}\right)^{M+1} \tag{3.39}$$

Substituting (3.39) into (3.35), the mean-squared error convergence is given by

$$\begin{aligned} \xi_n &= \sigma_d^2 \sum_{i=1}^{N} \frac{1}{N} \left(1 - \frac{1}{N}\right)^{n(M+1)} \\ &= \sigma_d^2 \left(1 - \frac{1}{N}\right)^{n(M+1)} \end{aligned} \tag{3.40}$$

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

$$\begin{aligned} \xi_{n,dB} &= 10\log_{10}\sigma_d^2 + 10n(M+1)\log_{10}\left(1 - \frac{1}{N}\right) \\ &\approx 10\log_{10}\sigma_d^2 - 4.343\frac{n(M+1)}{N} \end{aligned} \tag{3.41}$$

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 [18].

*Observation 6:* *NLMS is the special case of APA with M* = 0. If $M = 0$, then $\beta_i = p_i$ and difference equation (3.19), which describes the behavior of $\tilde{\lambda}_{n,i}$, becomes

$$\tilde{\lambda}_{n+1,i} = (1 - \alpha p_i)\tilde{\lambda}_{n,i} + \bar{\mu}^2 \xi^0 E\left(\frac{1}{r^2}\right) p_i. \tag{3.42}$$

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

$$\xi_n = \sigma_d^2 \sum_{i=1}^{N} (1 - \alpha p_i)^n p_i. \tag{3.43}$$

These results match the earlier results derived for NLMS under the same assumptions [4]. From Observation 5, 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 [20].

*A Special Comment for PRA.* The partial rank algorithm 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, defined in (3.14), become for PRA:

$$\tilde{\lambda}_{n+1,i} = \begin{cases} \tilde{\lambda}_{n,i}, & if\ ((n+1))_{M+1} \neq 0 \\ (1-\alpha\beta_i)\tilde{\lambda}_{n,i} + \bar{\mu}^2 \xi^0 E\left(\frac{1}{r^2}\right)\beta_i, & if\ ((n+1))_{M+1} = 0 \end{cases} \tag{3.44}$$

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

$$\xi_n = \sigma_d^2 \sum_{i=1}^{N} (1-\alpha\beta_i)^{\left\lfloor \frac{n}{M+1} \right\rfloor} p_i, \tag{3.45}$$

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

## 3.2 Verification Using Simulation

In this section, we demonstrate the validity of the analytical results presented in Section 3.1 and also 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 (3.30) 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, i.e. $\xi^0 = 0$, unless noted otherwise. The simulation results shown are obtained by ensemble averaging over 100 independent trials of the experiment.

Figure 3.2 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 (3.35), 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, while this is not true when $D$ = 1. This is an advantage of NLMS-OCF, which allows $D > 1$.

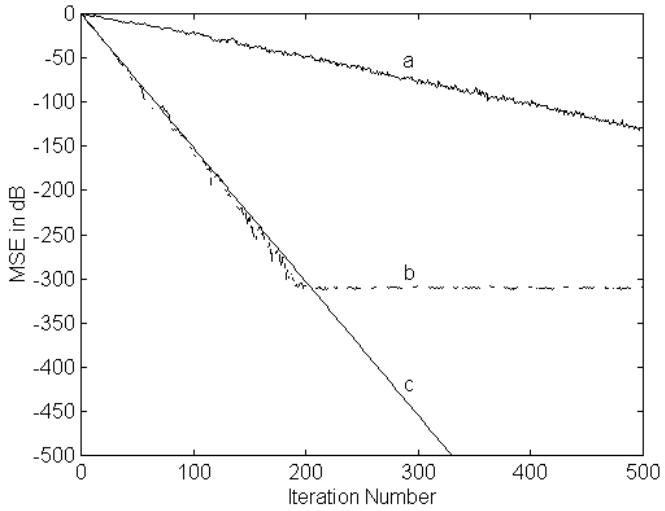


**Figure 3.2 Learning Curves of APA for White Input Using $\bar{\mu}$=1.0**
**(a) Simulated with *D*=1, (b) Simulated with *D*=32, and (c) Theoretical.**
**(Input: White Noise, System: FIR(31), $\xi^0 = 0$, and *M*=10)**

The results obtained using the reasonably colored signal as input are shown in Figure 3.3. 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.

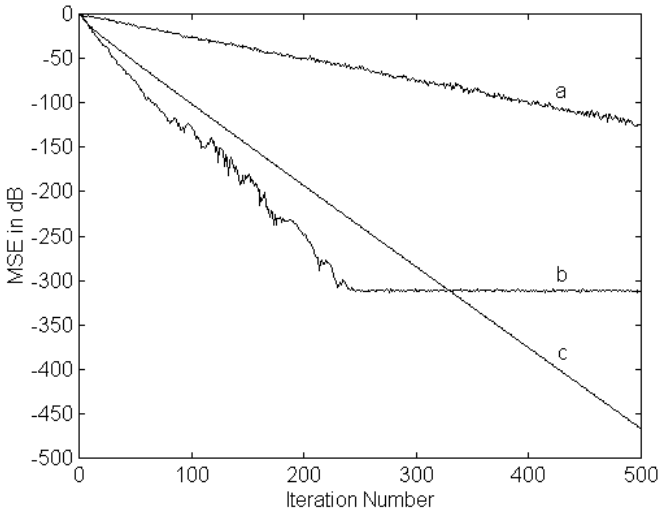


**Figure 3.3 Learning Curves of APA for Reasonably Colored Input Using $\bar{\mu}$=1.0**
**(a) Simulated with *D*=1, (b) Simulated with *D*=32, and (c) Theoretical.**
**(Input: AR(1), pole at 0.25, System: FIR(31), $\xi^0 = 0$, and *M*=10)**

Results, for the highly colored signal as input, similar to the results shown in Figures 3.2 and 3.3, are shown in Figure 3.4. 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. One would expect this behavior, since the highly correlated input violates the independence assumption more strongly than the other two inputs.

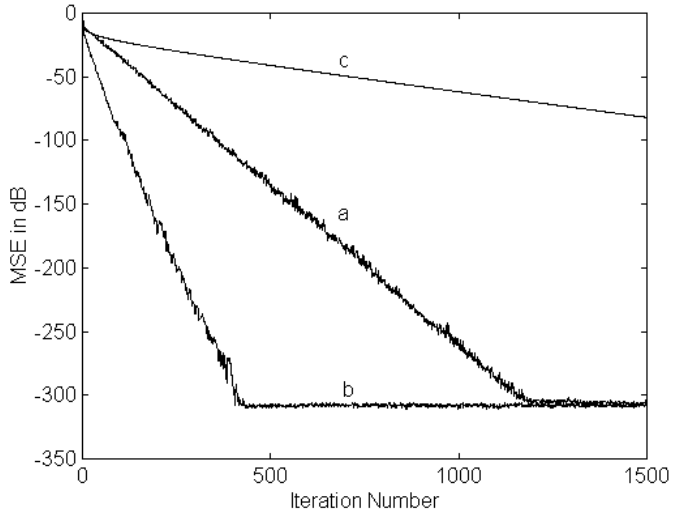


**Figure 3.4 Learning Curves of APA for Highly Colored Input Using $\bar{\mu}$=1.0**
**(a) Simulated with *D*=1, (b) Simulated with *D*=32, and (c) Theoretical.**
**(Input: AR(1), pole at 0.95, System: FIR(31), $\xi^0 = 0$, and *M*=10)**

From Figures 3.2 through 3.4 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 $k \le n$. The dependence of $\mathbf{w}_n$ on the past input vectors can also be reduced by using a smaller value for the step size. Due to 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 Figures 3.3 and 3.5, which are obtained using the reasonably colored signal. For 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. Also, note that the convergence rate is slower with $\bar{\mu}$ = 0.1 than with $\bar{\mu}$ = 1.

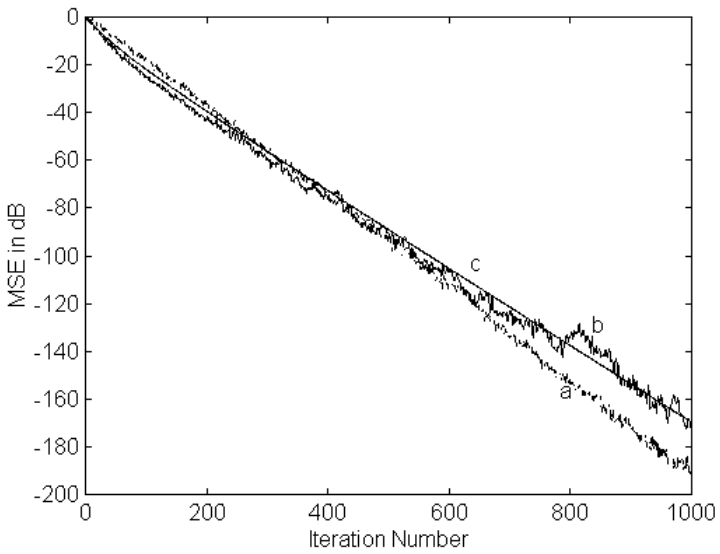


**Figure 3.5 Learning Curves of APA for Reasonably Colored Input Using $\bar{\mu}$=0.1**
**(a) Simulated with *D*=1, (b) Simulated with *D*=32, and (c) Theoretical.**
**(Input: AR(1), pole at 0.25, System: FIR(31), $\xi^0 = 0$, and *M*=10)**

The simulation results and theoretical results for the highly colored input signal are shown in Figure 3.6. Here also 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 Figures 3.2 through 3.4 the convergence for $D$ = 32 is faster than for $D$ = 1. The latter behavior is not universal, as the results in Figure 3.6 illustrate.

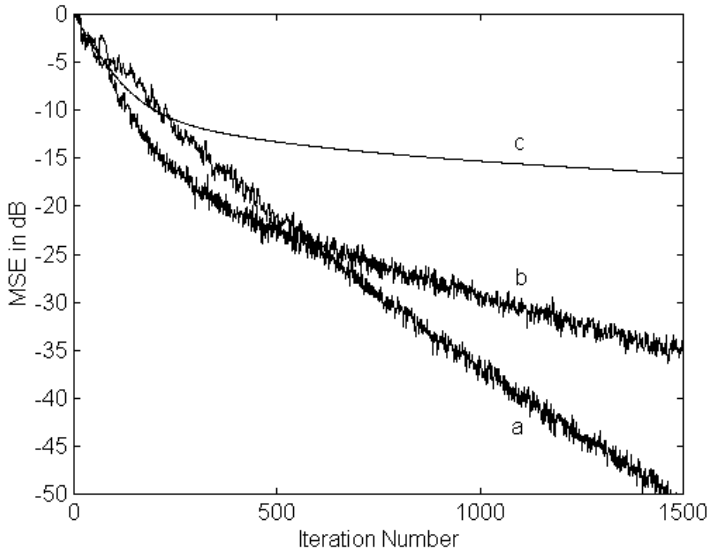


**Figure 3.6 Learning Curves of APA for Highly Colored Input Using $\bar{\mu}$=0.01 (a) Simulated with *D*=1, (b) Simulated with *D*=32, and (b) Theoretical. (Input: AR(1), pole at 0.95, System: FIR(31), $\xi^0 = 0$, and *M*=10)**

Figure 3.7 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 14000 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 4 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.

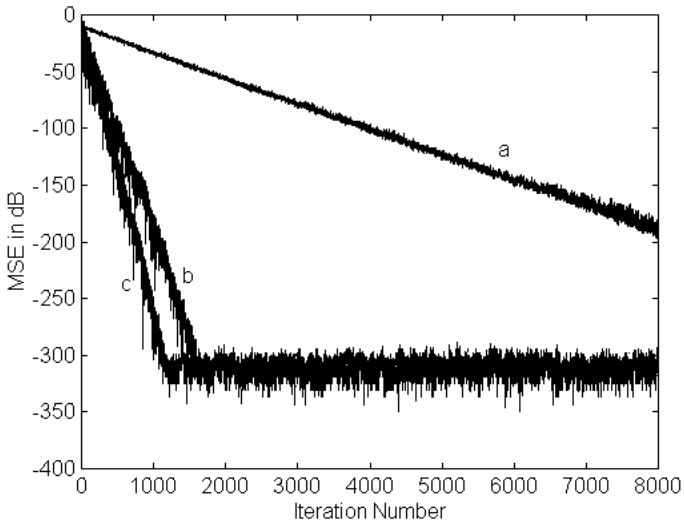


**Figure 3.7 Simulated Learning Curves of APA for Highly Colored Input – Various *M***
**(a) *M*=0 (NLMS), (b) *M*=2, and (c) *M*=8.**
**(Input: AR(1), pole at 0.95, System: FIR(31), $\xi^0 = 0$, $\bar{\mu}$ =1.0, and *D*=1)**

The simulation results with white noise input, for different values of *M*, as shown in Figure 3.8, corroborate Observation 5. While 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 [20] suggests that APA will not converge faster than NLMS when the input is white, which cannot be decorrelated any further by APA.

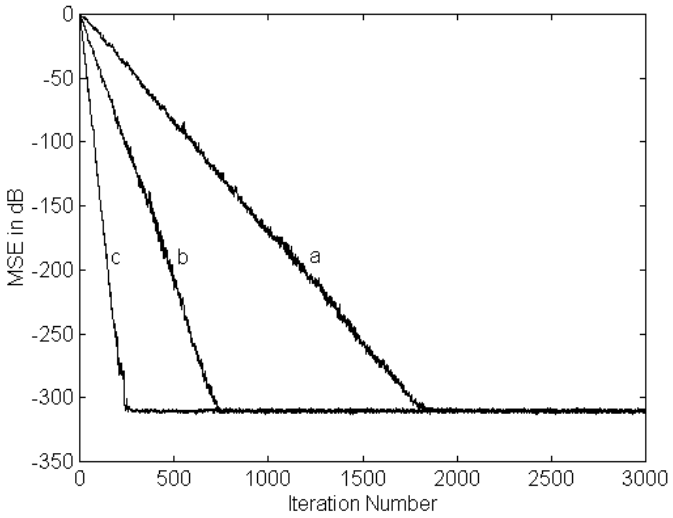


**Figure 3.8 Simulated Learning Curves of APA for White Input – Various *M***
**(a) *M*=0 (NLMS), (b) *M*=2, and (c) *M*=8.**
**(Input: White Noise, System: FIR(31), $\xi^0 = 0$, $\bar{\mu}$=1.0, and *D*=1)**

Observation 2 suggested that APA provides a way to improve the convergence rate without compromising on misadjustment. The following experiment corroborates this observation. Figure 3.9a 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). Figures 3.9b and 3.9c 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, while the resulting misadjustments are quite different. NLMS with $\bar{\mu}$ = 1 has a misadjustment $\mathcal{M}$ of 1.1164 while 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, while 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). 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.

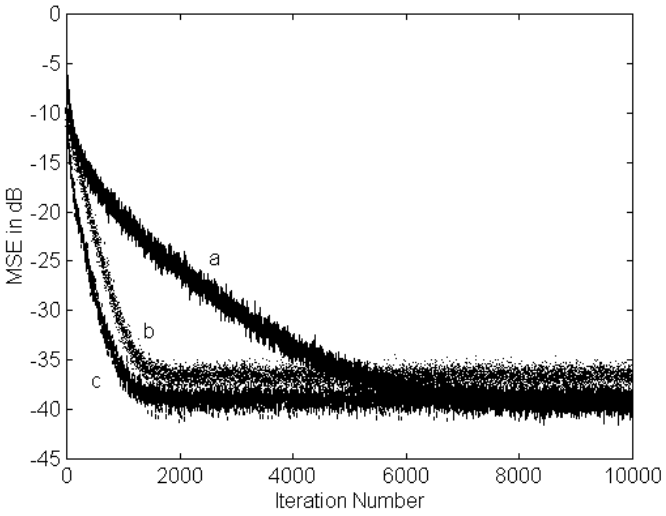


**Figure 3.9 Simulated Learning Curves of APA**
**Misadjustment/Convergence Rate Trade-Off:**
**(a) *M*=0 (NLMS) and $\bar{\mu}$=0.25, (b) *M*=0 (NLMS) and $\bar{\mu}$=1.0, (c) *M*=2 and $\bar{\mu}$=0.25.**
**(Input: AR(1), pole at 0.95, System: FIR(31), $\xi^0 = 10^{-4}$, and *D*=32)**

Figure 3.10 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, 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 2 that the misadjustment for APA does not depend on *M* holds as long as the data and parameters satisfy our assumptions.

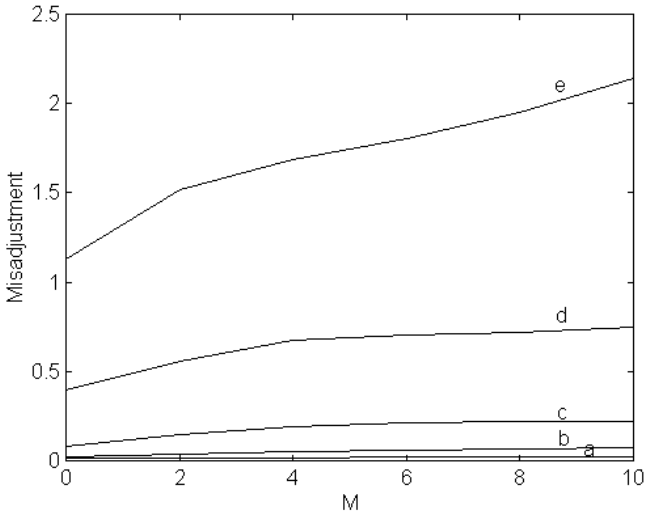


**Figure 3.10 Dependence of Misadjustment on Stepsize**
**(a) $\bar{\mu}$ =0.001, (b) $\bar{\mu}$ =0.01, (c) $\bar{\mu}$ =0.1, (d) $\bar{\mu}$ =0.5, and (e) $\bar{\mu}$ =1.0.**
**(Input: AR(1), pole at 0.95, System: FIR(31), $\xi^0 = 10^{-4}$, and *D*=32)**

## 3.3 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 and/or step-sizes tend towards satisfying the independence assumption the simulated results tend to the theoretical results, while 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.

While 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. 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. Simulation results corroborate our findings.