CONVERGENCE ANALYSIS OF THE NLMS ALGORITHM WITH M-INDEPENDENT INPUTS

Pascal Scalart

FRANCE TELECOM R&D, 2. Av. Pierre Marzin, 22307 Lannion Cedex, FRANCE {e-mail: pascal.scalart@rd.francetelecom.fr}

ABSTRACT

In most adaptive identification applications, a finite impulse response (FIR) filter is employed with coefficients that are computed using the normalized least mean square (NLMS) algorithm. In this paper, the convergence behavior of the NLMS algorithm is analyzed using a simple model of the input signal vectors. Explicit expressions of the learning curve and misadjustement are derived and compared with those previously established for the NLMS algorithm. Comparisons between theoretical and experimental results are given to validate our approach.

1. INTRODUCTION

In this paper, we consider direct adaptive identification with transversal adaptive filters, which is the usual framework in many practical applications. In direct identification, the unknown system is characterized by two observations, namely, the input signal x_t and the signal y_t available at the output of the unknown system. The convolution of the excitation signal x_t with the L coefficients of the impulse response of a FIR filter $\mathbf{H}_{t-1}^T$ produces an estimate $\hat{y}_t$, which is subtracted from y_t to give an estimation error $e_t = y_t - \mathbf{H}_{t-1}^T \mathbf{X}_t$.

The NLMS algorithm is the most popular algorithm for updating the impulse response of the FIR adaptive filter. It provides an efficient way to implement the optimal L-samples Wiener filter that minimizes, in a stochastic approximation sense, the mean-square value of the filtering error (MSE) according to

$$\mathbf{H}_t = \mathbf{H}_{t-1} + \mu e_t \left(\mathbf{X}_t^T \mathbf{X}_t\right)^{-1} \mathbf{X}_t . \tag{1}$$

Assuming that the optimal filter $\mathbf{H}^{\text{opt}}$ of the unknown system is a FIR filter of order L, *i.e.* $y_t = \left(\mathbf{H}^{\text{opt}}\right)^T \mathbf{X}_t + \varepsilon_t$, the error $\Delta\mathbf{H}_t = \mathbf{H}^{\text{opt}} - \mathbf{H}_t$ in the estimated filter coefficients at time t may be expressed from (1) as

$$\Delta\mathbf{H}_t = \left[\mathbf{I} - \mu \mathbf{X}_t \left(\mathbf{X}_t^T \mathbf{X}_t\right)^{-1} \mathbf{X}_t^T\right] \Delta\mathbf{H}_{t-1} - \mu \frac{\mathbf{X}_t \varepsilon_t}{\mathbf{X}_t^T \mathbf{X}_t} \tag{2}$$

In the noiseless case and with $\mu = 1$, the interpretation of (2) leads to the projection interpretation of the NLMS. In cases where $\mu \neq 1$, this operation is defined as a relaxed affine projection of vector $\Delta\mathbf{H}_{t-1}$ on a subspace completely specified by the matrix within the square brackets of (2).

In this paper, we present an analysis of the convergence behavior of the NLMS algorithm. In Section 2, we briefly review the original work reported in [1] under the classical independence assumption. We then extend the analysis of the NLMS algorithm to the case of M-Independent inputs. In Section 3, we discuss the cases of 2- and 3-Independence and present some remarks regarding the convergence domain and steady-state behavior of the NLMS algorithm. Finally, we compare in Section 4 simulated and theoretical learning curves to validate our approach.

2. PRELIMINARIES

2.1. Model of the Input Signal Vectors

In [1], a specific model for the input signal vectors $\{\mathbf{X}_t\}$ has been proposed to analyze the convergence behavior of the LMS and NLMS algorithms. The distribution of $\mathbf{X}_t$ is modeled as the product of two marginal distributions: the "radial" distribution (random variable r) and the "angular" distribution (random variable $s\mathbf{W}_t$) of $\mathbf{X}_t$. Such a model is based on the following assumptions:

(**A1**) The sequence of input vectors $\{\mathbf{X}_t\}$ is independent and identically distributed (i.i.d.);

(**A2**) The vector $\mathbf{X}_t$ is the product of three independent variables that are i.i.d., *i.e.* $\mathbf{X}_t = s\, r \mathbf{W}_t$

$$\text{with} \begin{cases} \Pr(s = \pm 1) = 0.5 \\ r \approx \|\mathbf{X}_t\| \ (\approx \text{ means "as the same distribution as"}) \\ \Pr(\mathbf{W}_t = \mathbf{V}_i) = p_i = \lambda_i / tr(\mathbf{R}) \end{cases} \tag{3}$$

where *tr* denotes *trace*, and $\{\mathbf{V}_i, i \in [1, L]\}$ represent the eigenvectors of the decomposition of the covariance matrix of the input sequence $\mathbf{R} = E\left[\mathbf{X}_t \mathbf{X}_t^T\right]$, *i.e.*,

$$\mathbf{R} = \mathbf{V}\Sigma\mathbf{V}^T = \sum_{i=1}^{L} \lambda_i \mathbf{V}_i \mathbf{V}_i^T . \tag{4}$$

It is satisfying to note that $E[\mathbf{X}_t] = 0$ and $E\left[\mathbf{X}\,\mathbf{X}^T\right] = \mathbf{R}$. By using this model in the NLMS coefficient update formula, it has been demonstrated in [1] a close correspondence between the theoretical learning curves of the NLMS algorithm and the simulations using experimental data. The following sub-section recalls some of the main results established in [1].

2.2. The NLMS learning curve

We will assume in the following that ε_t is i.i.d., with zero mean and variance σ^2, and is independent of $\{\mathbf{X}_t\}$. Furthermore, we shall assume that the optimal filter $\mathbf{H}^{\text{opt}}$ and the identification filter $\mathbf{H}_t$ have the same length of L taps. We can rewrite the *a priori* error $e_t = \varepsilon_t + \Delta\mathbf{H}_{t-1}^T \mathbf{X}_t$ and the mean-square error (MSE) as

$$\xi_t = E\left[e_t^2\right] = \sigma^2 + E\left[\mathbf{X}_t^T \Delta\mathbf{H}_{t-1} \Delta\mathbf{H}_{t-1}^T \mathbf{X}_t\right] \tag{5}$$

$$\xi_t = \sigma^2 + \sum_{i=1}^{L} \lambda_i \tilde{\lambda}_i (t-1) \quad (6)$$

where independence of the input sequence vectors $\{\mathbf{X}_t\}$ has been used, and where the diagonal elements $\tilde{\lambda}_i(t)$ are given by

$$\tilde{\lambda}_i(t) = \mathbf{V}_i^T \mathbf{Cov}_t \mathbf{V}_i . \quad (7)$$

with $\mathbf{Cov}_t = E\left[\Delta\mathbf{H}_t \Delta\mathbf{H}_t^T\right]$. Replacing $\Delta\mathbf{H}_t$ by the relation of (2), we get

$$\tilde{\lambda}_i(t) = \mu^2\sigma^2 p_i E\left[\frac{1}{r_t^2}\right] + \left[1-\mu(2-\mu)p_i\right]\tilde{\lambda}_i(t-1) . \quad (8)$$

Analyzing the steady-state behavior of the above equation and replacing $\tilde{\lambda}_i(t-1)$ in (6) by the asymptotic value $\lim_{t\to+\infty}\tilde{\lambda}_i(t)$, we get the steady-state MSE, or equivalently the misadjustment (with independence assumption) given by

$$M_{Ind} = \frac{\lim_{t\to+\infty}\xi_t - \sigma^2}{\sigma^2} = \frac{\mu}{2-\mu} tr(\mathbf{R}) E\left[\frac{1}{r^2}\right]. \quad (9)$$

3. M-INDEPENDENT INPUTS

In this section, we analyze the convergence behavior of the NLMS algorithm (i.e. learning curve and misadjustment) with a modified model for the input signal vectors $\{\mathbf{X}_t\}$.

3.1. New model of the input signal vectors

The NLMS analysis given in section 2 use the so-called independence assumption which specifies that the sequence of input vectors $\{\mathbf{X}_t\}$ is an i.i.d. sequence. In this section, a generalization of the model given by (3) is proposed and the joint probability between successive input signal vectors is introduced. In particular, the vector $\mathbf{X}_t$ is now modeled as an M-independent process, *i.e.* independence is assumed between the sequences $\{\mathbf{X}_t\}$ and $\{\mathbf{X}_{t-M}\}$. The input sequence vector is given by $\mathbf{X}_t = s\,r\mathbf{W}_t$ where variables s and r are assumed independent, and with the joint probability given by

$$\Pr\left(\mathbf{W}_t = \mathbf{V}_i, \cdots, \mathbf{W}_{t-M-1} = \mathbf{V}_p\right) = p_{\underbrace{i,\cdots,p}_{\text{M subscrips}}} \quad (10)$$

where $\sum_{j,\cdots,p} p_{ij\cdots p} = p_i = \lambda_i / tr(\mathbf{R})$.

The introduction of the joint probability between successive input vectors provides an analysis which is not restricted to the case of independence assumption which is clearly violated in practice since $\mathbf{X}_t$ and $\mathbf{X}_{t-1}$ have $L-1$ samples in common. In the following, we restricted our analysis to the 2- and 3-independent cases.

3.2. Learning curve with 2-independence assumption

Let us first consider the case of 2-independence assumption for the input signal vectors where $\Pr\left(\mathbf{W}_t = \mathbf{V}_i, \mathbf{W}_{t-1} = \mathbf{V}_j\right) = p_{ij}$. Inserting (2) in relation (5), we can re-write the MSE as

$$\xi_t = \sigma^2 + \mu^2\sigma^2 E\left[\mathbf{X}_t^T \frac{\mathbf{X}_{t-1}\mathbf{X}_{t-1}^T}{\left(\mathbf{X}_{t-1}^T\mathbf{X}_{t-1}\right)^2}\mathbf{X}_t\right] + E\left[\mathbf{X}_t^T\left(\mathbf{I}-\mu\frac{\mathbf{X}_{t-1}\mathbf{X}_{t-1}^T}{\mathbf{X}_{t-1}^T\mathbf{X}_{t-1}}\right)\Delta\mathbf{H}_{t-2}\Delta\mathbf{H}_{t-2}^T\left(\mathbf{I}-\mu\frac{\mathbf{X}_{t-1}\mathbf{X}_{t-1}^T}{\mathbf{X}_{t-1}^T\mathbf{X}_{t-1}}\right)\mathbf{X}_t\right] .$$

Since $\{\mathbf{X}_t\}$ and $\{\mathbf{X}_{t-2}\}$ are assumed independent, introduction of the probabilistic model (10) in the above equation leads to

$$\xi_t = \sigma^2 + \mu^2\sigma^2\sum_{i=1}^{L}\sum_{j=1}^{L} p_{ij}\mathbf{V}_i^T\mathbf{V}_j\mathbf{V}_j^T\mathbf{V}_i + E\left[r^2\right]\sum_{i=1}^{L}\sum_{j=1}^{L} p_{ij}\mathbf{V}_i^T\left(\mathbf{I}-\mu\mathbf{V}_j\mathbf{V}_j^T\right)\mathbf{Cov}_{t-2}\left(\mathbf{I}-\mu\mathbf{V}_j\mathbf{V}_j^T\right)\mathbf{V}_i$$

that is

$$\xi_t = \sigma^2 + \mu^2\sigma^2\sum_{i=1}^{L} p_{ii} + E\left[r^2\right]\sum_{i=1}^{L}\left[p_i - \mu(2-\mu)p_{ii}\right]\tilde{\lambda}_i(t-2) . \quad (11)$$

Replacing twice $\Delta\mathbf{H}_t$ in (7) by the equality of (2), we get

$$\tilde{\lambda}_i(t) = \mu^2\sigma^2 E\left[\mathbf{V}_i^T\frac{\mathbf{X}_t\mathbf{X}_t^T}{\left(\mathbf{X}_t^T\mathbf{X}_t\right)^2}\mathbf{V}_i\right] + \mu^2\sigma^2 E\left[\mathbf{V}_i^T\left(\mathbf{I}-\mu\frac{\mathbf{X}_t\mathbf{X}_t^T}{\mathbf{X}_t^T\mathbf{X}_t}\right)\frac{\mathbf{X}_{t-1}\mathbf{X}_{t-1}^T}{\left(\mathbf{X}_{t-1}^T\mathbf{X}_{t-1}\right)^2}\left(\mathbf{I}-\mu\frac{\mathbf{X}_t\mathbf{X}_t^T}{\mathbf{X}_t^T\mathbf{X}_t}\right)\mathbf{V}_i\right] + E\left[\mathbf{V}_i^T\prod_{n=0}^{1}\left(\mathbf{I}-\mu\frac{\mathbf{X}_{t-n}\mathbf{X}_{t-n}^T}{\mathbf{X}_{t-n}^T\mathbf{X}_{t-n}}\right)\Delta\mathbf{H}_{t-2}\Delta\mathbf{H}_{t-2}^T\prod_{n=0}^{1}\left(\mathbf{I}-\mu\frac{\mathbf{X}_{t-n}\mathbf{X}_{t-n}^T}{\mathbf{X}_{t-n}^T\mathbf{X}_{t-n}}\right)\mathbf{V}_i\right]$$

which is

$$\tilde{\lambda}_i(t) = \mu^2\sigma^2 E\left[\frac{1}{r^2}\right]\left[2p_i - \mu(2-\mu)p_{ii}\right] + \left[1-\mu(2-\mu)\left[2p_i-\mu(2-\mu)p_{ii}\right]\right]\tilde{\lambda}_i(t-2) . \quad (12)$$

After some simplifications, the misadjustment is given by

$$M_{2-Ind} = M_{Ind} + \mu^2\sum_{i=1}^{L} p_{ii}\left(1 - tr(\mathbf{R})E\left[\frac{1}{r^2}\right]\right). \quad (13)$$

3.3. Learning Curve with 3-independence assumption

Following directly from the results given in the previous sub-section, we can write for the 3-independent case:

$$\xi_t = \sigma^2 + \mu^2\sigma^2 E\left[\mathbf{X}_t^T\frac{\mathbf{X}_{t-1}\mathbf{X}_{t-1}^T}{\left(\mathbf{X}_{t-1}^T\mathbf{X}_{t-1}\right)^2}\mathbf{X}_t\right] + \mu^2\sigma^2 E\left[\mathbf{X}_t^T\left(\mathbf{I}-\mu\frac{\mathbf{X}_{t-1}\mathbf{X}_{t-1}^T}{\mathbf{X}_{t-1}^T\mathbf{X}_{t-1}}\right)\frac{\mathbf{X}_{t-2}\mathbf{X}_{t-2}^T}{\left(\mathbf{X}_{t-2}^T\mathbf{X}_{t-2}\right)^2}\left(\mathbf{I}-\mu\frac{\mathbf{X}_{t-1}\mathbf{X}_{t-1}^T}{\mathbf{X}_{t-1}^T\mathbf{X}_{t-1}}\right)\mathbf{X}_t\right] + E\left[\mathbf{X}_t^T\prod_{n=1}^{2}\left(\mathbf{I}-\mu\frac{\mathbf{X}_{t-n}\mathbf{X}_{t-n}^T}{\mathbf{X}_{t-n}^T\mathbf{X}_{t-n}}\right)\Delta\mathbf{H}_{t-3}\Delta\mathbf{H}_{t-3}^T\prod_{n=1}^{2}\left(\mathbf{I}-\mu\frac{\mathbf{X}_{t-n}\mathbf{X}_{t-n}^T}{\mathbf{X}_{t-n}^T\mathbf{X}_{t-n}}\right)\mathbf{X}_t\right]$$

Consequently, inserting the probabilistic model for the 3-independence case, we obtain

$$\xi_t = \sigma^2 + \mu^2\sigma^2\sum_{i=1}^{L} q_i(\mu) + E\left[r^2\right]\sum_{i=1}^{L}\left[p_i - \mu(2-\mu)q_i(\mu)\right]\tilde{\lambda}_i(t-3)$$

where : $$q_i(\mu) = p_{ii} + \sum_{j=1}^{L} p_{iji} - \mu(2-\mu)\, p_{iii}\,. \qquad (14)$$

Proceeding the same way as we did for (12) and replacing three times $\mathbf{\Delta H}_t$ in (7) by the equality of (2), we find that the diagonal elements $\tilde{\lambda}_i(t)$ are given by

$$\tilde{\lambda}_i(t) = \mu^2\sigma^2 E\left[\frac{1}{r^2}\right] r_i(\mu) + \left[1-\mu(2-\mu) r_i(\mu)\right]\tilde{\lambda}_i(t-3) \qquad (15)$$

with

$$r_i(\mu) = 3p_i - \mu(2-\mu)\left[2p_{ii} - \mu(2-\mu)\, p_{iii}\right] - \mu\left[p_{ii} + (1-\mu)\sum_{k=1}^{L} p_{iki}\right]\,.$$

Analyzing the steady-state behavior of (15), it can be shown that the misadjustment is given by

$$M_{3-Ind} = M_{Ind} + \mu^2 \sum_{i=1}^{L} q_i(\mu)\left(1 - tr(\mathbf{R})\, E\left[\frac{1}{r^2}\right]\right). \qquad (16)$$

3.4. Convergence and Misadjustement

Based on the analysis of the previous sections [from (8), (12), and (15)], it follows that a sufficient condition for the convergence of the NLMS algorithm is given by $0 < \mu < 2$, and the fastest convergence occurs for $\mu = 1$. Moreover, inserting in (9), (13), and (16) the following approximation (from [1]),

$$E\left[\frac{1}{r^2}\right] \approx \frac{1}{tr(\mathbf{R})}\left(1 - \frac{\nu_x - 1}{L}\right)^{-1}$$

which holds if the kurtosis of the input signal satisfies $\nu_x = E\left[x_t^{\,4}\right] \big/ \sigma_x^4 << L$, we can see that

$$M_{3-Ind} \le M_{2-Ind} \le M_{Ind} \qquad (17)$$

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, we present computer simulations to illustrate the usefulness of the proposed approach. For these simulations, we have used the same parameters than those presented in [1]: the input signal is a gaussian first-order autoregressive (AR) process with pole $\alpha = 0$ and $\alpha = 0.9$, the step-size of the algorithm is set to unity, the filter length L is equal to 20, the initial value of the filter coefficient vector is $\mathbf{H}_{-1} = \mathbf{0}$, and the optimal filter is chosen to ensure that the initial diagonal elements $\tilde{\lambda}_i(-1)$ are equal, *i.e.* the optimal filter has components of equal magnitude along all eigenvectors of $\mathbf{R}$. In the following sub-sections, we compare simulations of the NLMS algorithm and the learning curves predicted by the theory from (6), (11), and (14).

4.1. Source Probability Estimate

In practice, the theoretical values of the joint probability of the input signal vectors are not known *a priori*. However, without affecting the performance, we can estimate this probability from Monte-Carlo simulations. For this purpose, we first select for each vector $\mathbf{X}_t$ the nearest eigenvector $\hat{\mathbf{U}}_t$ (in the MSE sense), that is

$$\hat{\mathbf{U}}_t = \underset{\mathbf{U}\in[\mathbf{V}_1,\mathbf{V}_2,\cdots\mathbf{V}_L]}{\operatorname{argmin}} \left\{ \left\| \frac{\mathbf{X}_t}{\|\mathbf{X}_t\|} - \mathbf{U} \right\|^2, \left\| \frac{\mathbf{X}_t}{\|\mathbf{X}_t\|} + \mathbf{U} \right\|^2 \right\}. \qquad (18)$$

Then, we estimate $\Pr\left(\mathbf{W}_t = \hat{\mathbf{U}}_t, \mathbf{W}_{t-1} = \hat{\mathbf{U}}_{t-1}, \cdots, \mathbf{W}_{t-p} = \hat{\mathbf{U}}_{t-p}\right)$ as the ratio between the number of events and the total number of experiments (85 millions in our case). For a confidence interval of 95%, this value provides a relative precision of 0.002 and 0.66 for the estimation of a probability of 10^{-2} and 10^{-7} respectively. In the two-dimensional case, the experimental joint probability is given in Fig. 1 [plot of $\log_{10}(p_{ij})$] for an AR(1) process with pole $\alpha = 0$ and $\alpha = 0.9$ respectively. To demonstrate the accuracy of this procedure, we found that the difference between theoretical and experimental marginal probabilities was inferior to 0.1% (relative to the theoretical ones) for the white noise case ($\alpha = 0$), and continuously increases from 0.3% (maximum eigenvalue) to 98% (minimum eigenvalue) for the colored case ($\alpha = 0.9$).

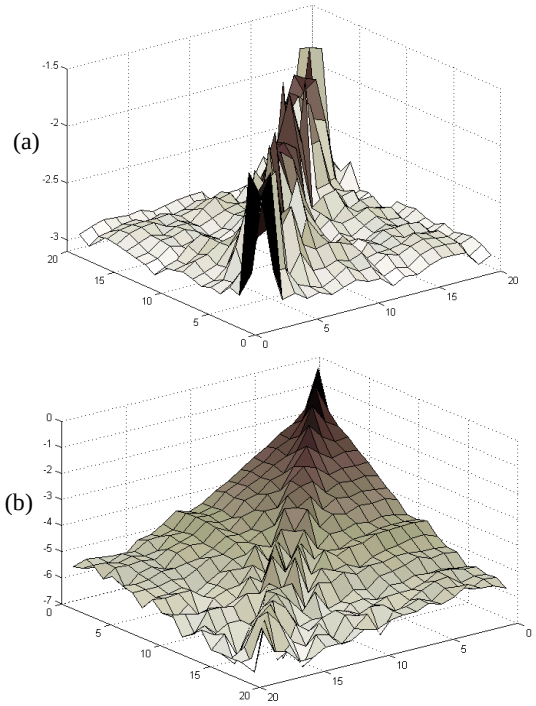

Fig. 1. Experimental two-dimensional probability for gaussian AR(1) input process with (a) pole $\alpha = 0$, and (b) pole $\alpha = 0.9$.

4.2. Learning curve

By inserting the previous joint probability estimates in relations (6), (11), and (14), we compare in noiseless case (see Fig. 2) Monte Carlo (100 runs are averaged) simulations of the NLMS algorithm (solid line) and the learning curves predicted from the theory (bold lines) for a gaussian AR(1) input process with pole $\alpha = 0$ and $\alpha = 0.9$ respectively. Also shown in Fig. 2 are the simulation results of the NLMS algorithm with independent regression vectors $\mathbf{X}_k$ (dot gray curve). On this short time span, we can notice the capability of the theoretical learning curves to closely predict the behavior of the simulation results. In the white noise case, note also that the theoretical curves with independence

assumption seem to correspond more closely to the simulated learning curves than those predicted by the theory with 2- or 3-independence assumptions.

We have also considered in Fig. 2.(b) the curves resulting from our theory but with the independence assumption, i.e.

$$\Pr\left(\mathbf{W}_l = \mathbf{V}_i, \cdots, \mathbf{W}_{l-M-1} = \mathbf{V}_p\right) = \Pr\left(\mathbf{W}_l = \mathbf{V}_i\right) \cdots \Pr\left(\mathbf{W}_{l-M-1} = \mathbf{V}_p\right)$$

For the white noise and colored cases, one can easily observe that the three curves provided by the theory give approximately the same results and coincides with the simulation results of the NLMS algorithm with independent regression vectors.

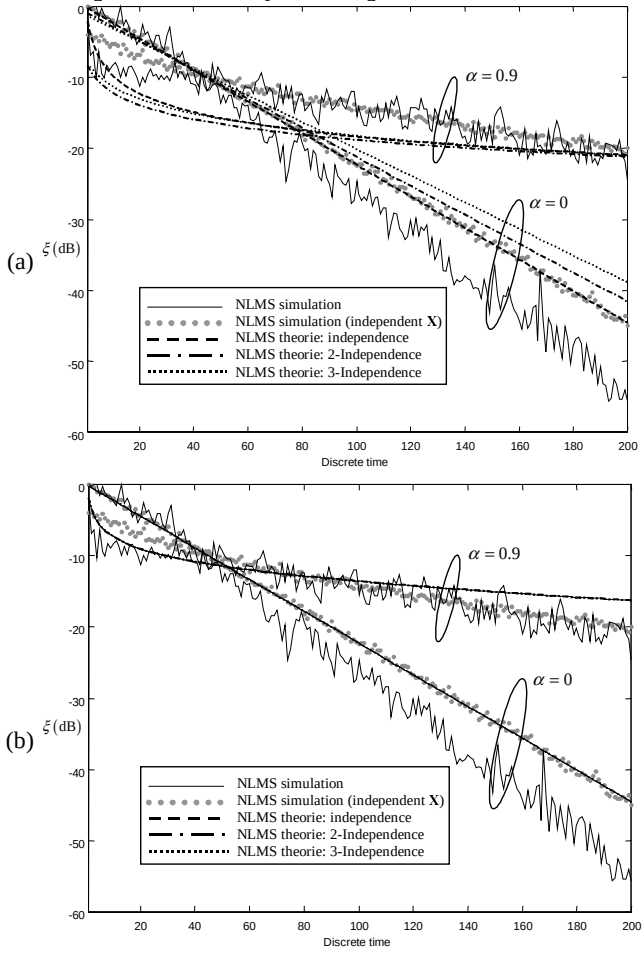

Fig. 2. Simulations of the NLMS algorithm compared with the theoretical learning curves with (a) experimental probabilities and (b) theoretical probabilities (independent assumption).

4.3. Misadjustement

In this simulation, the experimental values of the steady-state MSE are compared to those predicted by the theory, or equivalently the misadjustment computed from (9), (13), and (16) and using the experimental joint probabilities previously described in sub-section 4.1. These comparisons are given in Fig. 3 for the white noise and colored noise cases. The experimental results correspond to Monte-Carlo simulations (100 runs - SNR of 60 dB) of the NLMS algorithm for Gaussian AR(1) input process.

In the white noise case, Fig. 3.(a) shows the theoretical to experimental misadjustment ratio (in dB) given by

$$10\log_{10}\left(M_{\text{theo.}}/M_{\text{exp.}}\right) = 10\log_{10}\left(\xi^{\infty}_{\text{theo.}} - \sigma^2\right) - 10\log_{10}\left(\xi^{\infty}_{\text{exp.}} - \sigma^2\right)$$

It is easily observed that the experimental misadjustement coincides with the theoretical misadjustment predicted from (9), (13), and (16) with the ordering established in (17). Fig. 3.(b) shows that for highly correlated signals, the theoretical misadjustment computed from (9) with independent $\{\mathbf{X}_k\}$ exhibits a large bias whatever step-size is used. On the other hand, results computed for the 3- or 2-independence cases provide an accurate estimate of the experimental misadjustment, especially for large values of the step size.

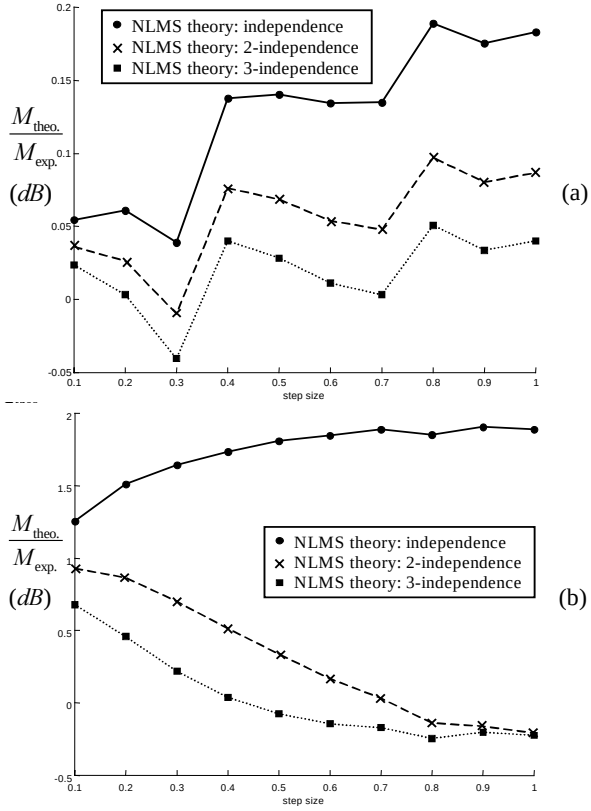

Fig. 3. Theoretical to experimental misajustement ratio (in dB) for gaussian AR(1) input process with pole (a) $\alpha = 0$ and (b) $\alpha = 0.9$.

5. SUMMARY

In this article, a theoretical analysis of the convergence and modal behavior of the learning curve of the NLMS algorithm is provided. By introducing the M-independence assumption for the input signal vectors in a specific distribution model, we have shown a close correspondence between the simulated and theoretical learning curves. Some interesting results are also given regarding analytic expressions for the misadjustment of the NLMS with M-independence assumption which closely predict the behavior of the experimental results for gaussian AR process.

6. REFERENCE

[1] D. T. M. SLOCK, "On the convergence behavior of the LMS and the normalized LMS algorithms," *in Proc. IEEE Trans. on Signal Proc.*, vol. 41, no. 9, pp. 2811-2825, Sept.1993.