STOCHASTIC GRADIENT ADAPTIVE FILTERS WITH GRADIENT ADAPTIVE STEP SIZES

V. John Mathews and Zhenhua Xie

Department of Electrical Engineering
University of Utah
Salt Lake City, Utah 84112

ABSTRACT

This paper presents two adaptive step-size gradient adaptive filters. The step sizes are changed using a gradient descent algorithm designed to minimize the squared estimation error. The first algorithm uses the same step-size sequence for all the filter coefficients whereas the second algorithm uses different step-size sequences for different adaptive filter coefficients. An analytical performance analysis of the first algorithm is also presented in the paper. Analyses and experiments indicate that (1) the algorithms have fast convergence rates and small misadjustment errors; and (2) in nonstationary environments, the algorithms tend to adjust the step sizes so as to give close to the best possible performance. Several simulation examples demonstrating the good properties of the adaptive filters are also presented in the paper.

I. INTRODUCTION

Stochastic gradient adaptive filters are extremely popular because of their inherent simplicity. However, they suffer from relatively slow and data-dependent convergence behavior. It is well known that the performance of stochastic gradient methods is adversely affected by high eigenvalue spreads of the autocorrelation matrix of the input vector.

Traditional approaches for improving the speed of convergence of the gradient adaptive filters have been to employ time-varying convergence parameters [1-5]. The idea is to somehow sense how far away the adaptive filter coefficients are from the optimal filter coefficients and use convergence parameters that are small when adaptive filter coefficients are close to the optimal values and use large convergence parameters otherwise. The approach is heuristically sound and has resulted in several *ad hoc* techniques, where the selection of the convergence parameter is based on the magnitude of the estimation error [3], polarity of the successive samples of the estimation error [4], measurement of the cross correlation of the estimation error with input data [1, 2], and so on. Experimentation with these techniques has shown that their performance is highly dependent on the selection of certain parameters in the algorithm and furthermore, the optimal choice of these parameters are highly data dependent. This fact has severely limited the usefulness of such algorithms in practical applications. Recently, Michael, *et al.* [5] have proposed methods for selecting the convergence parameters that would give the fastest speed of convergence. Unfortunately, their choices of the convergence parameters will also result in fairly large steady-state excess mean squared estimation error.

The objective of this paper is to present two stochastic gradient adaptive filtering algorithms that overcome the limitations of the methods discussed above. The idea that we will employ is to change the time-varying convergence parameters $\mu(n)$ in such a way that the change is proportional to the negative of the gradient of the squared estimation error with respect to the convergence parameter. This approach results in the following two algorithms.

This work was supported in part by NSF Grant # MIP 8708970.

Algorithm 1 (Single Convergence Sequence $\mu(n)$)

$$e(n) = d(n) - H^T(n)X(n) \tag{1}$$

$$\mu(n) = \mu(n-1) - \frac{\rho}{2}\frac{\partial}{\partial\mu(n-1)}e^2(n) \tag{2a}$$

$$= \mu(n-1) + \rho e(n)e(n-1)X^T(n-1)X(n) \tag{2b}$$

and

$$H(n+1) = H(n) - \frac{\mu(n)}{2}\frac{\partial e^2(n)}{\partial H(n)} \tag{3a}$$

$$= H(n) + \mu(n)e(n)X(n) \tag{3b}$$

In the above equations, ρ is a small positive constant, $d(n)$ is the desired response signal of the adaptive filter, $X(n)$ is the input vector to the adaptive filter, and $H(n)$ is the vector of adaptive filter coefficients, all at time n.

Algorithm 2 (Individual Convergence Sequence $\mu_i(n)$ for Each Coefficient

$$e(n) = d(n) - H^T(n)X(n) \tag{4}$$

$$\mu_i(n) = \mu_i(n-1) + \rho e(n)e(n-1)x_i(n)x_i(n-1) \tag{5}$$

and

$$h_i(n+1) = h_i(n) - \frac{\mu_i(n)}{2}\frac{\partial e^2(n)}{\partial h_i(n)} = h_i(n) + \mu_i(n)e(n)x_i(n). \tag{6}$$

Here, $h_i(n)$ and $x_i(n)$ are the i-th elements of $H(n)$ and $X(n)$, respectively.

Algorithm 1 was originally proposed by Sin and Lee [6]. We now present a convergence analysis for this algorithm.

II. PERFORMANCE ANALYSIS OF ALGORITHM 1

For the performance analysis, we will assume that the adaptive filter structure is that of an N-point FIR filter, and the input vector $X(n)$ is obtained as a vector formed by the most recent N samples of the input sequence x(n), i.e.,

$$X(n) = [x(n), x(n-1), ..., x(n-N+1)]^T \tag{7}$$

where $(\cdot)^T$ denotes the matrix transpose of $(\cdot)$. Let $H_{opt}(n)$ denote the optimal coefficient vector (in the minimum mean-squared estimation error sense) for estimating the desired response signal $d(n)$ using $X(n)$. We will assume that $H_{opt}(n)$ is time varying, and that the time variations are caused by a random disturbance of the optional coefficient process. Thus, the behavior of the optional coefficient process can be modeled as

$$H_{opt}(n) = H_{opt}(n-1) + C(n-1) \tag{8}$$

where $C(n-1)$ is the disturbance process that is a zero mean and white vector process with covariance matrix $\sigma_c^2 I$.

CH2847-2/90/0000-1385 $1.00 © 1990 IEEE

In order to make the analysis tractable, we will make use of the following assumptions and approximations.

i) X(n), d(n) are jointly Gaussian and zero mean random processes. X(n) is a stationary process. Moreover, {X(n), d(n)} is uncorrelated with {X(k), d(k)} if $n \neq k$. This is the commonly employed independence assumption and is seldom true in practice. However, analyses employing this assumption have produced reliable design rules in the past.

ii) Let
$$d(n) = X^T(n)H_{opt}(n) + \zeta(n) \tag{9}$$

where $\zeta(n)$ corresponds to the optimal estimation error process. We will assume that the triplet {X(n), C(n), $\zeta(n)$} are statistically independent random processes.

iii) We will assume that the convergence parameter $\mu(n)$ is statistically independent of X(n) and e(n). While this is never true, experiments have indicated that the approximations that $\mu(n)$ and $\mu^2(n)$ are uncorrelated with X(n) and e(n) is reasonably accurate for small values of ρ and relatively white input signals Note that the condition under which the above approximation is accurate is when the statistical fluctuations of $\mu(n)$ are small when compared with that of X(n) and e(n). This condition is, in general, satisfied for small values of ρ.

iv) We will use the approximation that the statistical expectation of $e^2(n)X(n)X^T(n)$ conditioned on the coefficient vector H(n) is the same as the unconditional expectation, i.e.,
$$E\{e^2(n)X(n)X^T(n)|H(n)\} \approx E\{e^2(n)X(n)X^T(n)\}. \tag{10}$$

This approximation has been successfully employed for performance analysis of adaptive filters equipped with the sign algorithm [7].

Mean Behavior of the Weight Vector

Let
$$V(n) = H(n) - H_{opt}(n) \tag{11}$$

denote the coefficient misalignment vector at time n. Then,
$$e(n) = \zeta(n) - V^T(n)X(n). \tag{12}$$

Substituting (9), (11) and (12) into (3b), we can easily show that
$$V(n+1) = (I - \mu(n)X(n)X^T(n))V(n) + \mu(n)X(n)\zeta(n) - C(n). \tag{13}$$

It is straightforward using the independence assumption and the uncorrelatedness of $\mu(n)$ with X(n) and e(n) to show that
$$E\{V(n+1)\} = (I - E\{\mu(n)\}R)E\{V(n)\} \tag{14}$$

where R is the autocorrelation matrix of the input vector X(n).

Mean Squared Behavior of the Weight Vector

Let
$$K(n) = E\{V(n)V^T(n)\} \tag{15}$$

denote a second moment matrix of the misalignment vector. Multiplying both sides of Eq. (13) with their respective transposes, we get the following equation:
$$\begin{aligned} V(n+1)V^T(n+1) = {} & (I - \mu(n)X(n)X^T(n))V(n)V^T(n)(I - \mu(n)X(n)X^T(n)) \\ & + \mu^2(n)\zeta^2(n)X(n)X^T(n) + C(n)C^T(n) \\ & + g(\mu(n), X(n), V(n), \zeta(n), C(n)), \end{aligned} \tag{16}$$

where $g(\mu(n), X(n), V(n), \zeta(n), C(n))$ corresponds to the sum of the six terms that are explicitly not listed in the expansion. Under our assumptions, the mean value of these six terms are all zero matrices. Combining usual analysis techniques for Gaussian input signals [8] with the assumption that $\mu(n)$ and $\mu^2(n)$ are uncorrelated with the data while taking the statistical expectation of (16) will result in the following evolution equation for the second moment matrix of the coefficient misalignment vector:
$$\begin{aligned} K(n+1) = {} & K(n) + E\{\mu(n)\}(RK(n) + K(n)R) \\ & + E\{\mu^2(n)\}(2RK(n)R + R\sigma_e^2(n)) + \sigma_c^2 I, \end{aligned} \tag{17}$$

where
$$\sigma_e^2(n) = \xi_{min} + \text{tr } RK(n) \tag{18}$$

and
$$\xi_{min} = E\{\zeta^2(n)\} \tag{19}$$

is the minimum value of the mean-squared estimation error and tr{(•)} denotes the trace of the matrix (•).

The mean and mean-squared behavior of $\mu(n)$ can be shown to follow the following nonlinear difference equations.
$$\begin{aligned} E\{\mu(n)\} = {} & E\{\mu(n-1)\}\,(1 - \rho\{\sigma_e^2(n-1)\text{tr}(R^2) + 2\text{tr}(R^3K(n-1))\}) \\ & + \rho\text{tr}(R^2K(n-1)) \end{aligned} \tag{20}$$

and
$$\begin{aligned} E\{\mu^2(n)\} = {} & E\{\mu^2(n-1)\}(1 - 2\rho\{\sigma_e^2(n-1)\text{tr}(R^2) + 2\text{tr}(R^3K(n-1)\}) \\ & + 2\rho E\{\mu(n-1)\}\text{tr}(R^2K(n-1)) \\ & + \rho^2\text{tr}(\{2R^2K(n) + \sigma_e^2(n)R\}\,\{2R^2K(n-1) + \sigma_e^2(n-1)R\}). \end{aligned} \tag{21}$$

Details are omitted because of space limitations.

Equations (17)-(21), completely characterize the mean-squared behavior of the coefficient misalignment vector. Deriving conditions on ρ for convergence of the evolution equations appears to be a very difficult task. However, we can guarantee convergence of K(n) by restricting $\mu(n)$ to be such that it always stays within the range that would ensure convergence. A sufficient, but not necessary, condition on $\mu(n)$ to ensure mean-squared convergence of the adaptive filter is [8]
$$0 < \mu(n) < \frac{2}{3\text{tr}\{R\}}. \tag{22}$$

If $\mu(n)$ falls outside this range, one can bring it inside the range by setting $\mu(n)$ to the closest of 0 and $2/3\text{tr}\{R\}$. Assuming that the system of evolution equations converges, we now proceed to study the steady-state behavior of the adaptive filter.

Steady-State Properties of the Adaptive Filter

Let $\overline{\mu_\infty}$, $\overline{\mu_\infty^2}$, $\sigma_e^2(\infty)$, and K_∞ represent the steady-state values of $E\{\mu(n)\}$, $E\{\mu^2(n)\}$, $\sigma_e^2(n)$, and K(n), respectively. Substituting these values for their counterparts in equations (17), (18), (20), and (21), will yield the following steady-state characterization of the adaptive filter behavior
$$\overline{\mu_\infty} = \overline{\mu_\infty}\,(1 - \rho\{\sigma_e^2(\infty)\text{tr}(R^2) + 2\text{tr}(R^3K_\infty)\}) + \rho\text{tr}(R^2K_\infty), \tag{23}$$

$$\begin{aligned} \overline{\mu_\infty^2} = {} & \overline{\mu_\infty^2}\,(1 - 2\rho\{\sigma_e^2(\infty)\text{tr}(R^2) + 2\text{tr}(R^3K_\infty)\}) \\ & + 2\rho\overline{\mu_\infty}\,\text{tr}\,R^2K_\infty + \rho^2\text{tr}(\{2R^2K_\infty + \sigma_e^2(\infty)R\}^2), \end{aligned} \tag{24}$$

$$\sigma_e^2(\infty) = \xi_{min} + \text{tr}(RK_\infty) \tag{25}$$

and
$$K_\infty = K_\infty + \overline{\mu_\infty}(RK_\infty + K_\infty R) + \overline{\mu_\infty^2}(2RK_\infty R + \sigma_e^2(\infty)R) + \sigma_c^2 I. \tag{26}$$

Solutions for $\overline{\mu_\infty}$ and $\overline{\mu_\infty^2}$ can be easily obtained in terms of $\sigma_e^2(\infty)$ and K_∞ as
$$\overline{\mu_\infty} = \frac{\text{tr}\{R^2K_\infty\}}{\sigma_e^2(\infty)\text{tr}\{R^2\} + 2\text{tr}\{R^3K_\infty\}} \tag{27}$$

and
$$\overline{\mu_\infty^2} = (\overline{\mu_\infty})^2 + \frac{\rho}{2}\frac{\text{tr}\{(2R^2K_\infty + \sigma_e^2(\infty)R)^2\}}{\sigma_e^2(\infty)\text{tr}\{R^2\} + 2\text{tr}\{R^3K_\infty\}}. \tag{28}$$

Under several simplifying approximations, we can also show that

$$K_\infty \approx \frac{\xi_{min}}{2}\frac{\overline{\mu_\infty^2}}{\overline{\mu_\infty}} I + \frac{1}{2\overline{\mu_\infty}} \sigma_\epsilon^2 R^{-1}. \quad (29)$$

Equations (27) - (29) describe the steady-state behavior of the adaptive filter. Unfortunately, they are highly coupled in the sense that each steady-state parameter depends on several others. Obtaining closed-form expressions for each parameter independently of the others appears difficult and one has to resort to numerical solution of the above three sets of simultaneous, nonlinear equations.

III. EXPERIMENTAL RESULTS

In this section, we will present the results of several experiments that demonstrate the good properties of the algorithms presented in this paper and also verify some of the analytical results of the last section. All the results presented are averages of 50 independent runs.

Simulation Example #1 We consider the problem of identifying a 5-point FIR filter with impulse response sequence

$$\{h(n);\ n = 0, 1, 2, 3, 4\} = \{0.2, 0.6, 1.0, 0.6, 0.2\} \quad (30)$$

from measurements of the input signal and the output signal that is corrupted by observation noise. The input signal (in all examples) was a pseudorandom zero mean and white Gaussian signal with unit variance. The observations noise was zero mean, white and Gaussian with variance 0.01 and was independent of the input signal. In Figure 1, we have plotted the trace of the second moment matrix of the misalignment vector (mean-squared norm of the coefficient misalignment vector) as a function of time. For the experiments, the coefficients of the adaptive filter were all set to zero initially. The other parameters were $\rho = 0.001$ for algorithm 1 and 0.005 for algorithm 2; and $\mu(0) = 0.08$. (All the parameters as well as the input signal and observation noise statistics are the same for all experiments.) Curves A and B are the theoretical and empirical curves, respectively, for algorithm 1. Curve C is the experimental results for algorithm 2. From the figure, we can see that the performance of both the algorithms are similar for this application. Furthermore, the analytical results show close agreement with the empirical results. In Figure 2, we have plotted the performance measure for the variation of the LMS algorithm [5] guaranteed to achieve the fastest rate of convergence against that of algorithm 1.. We can see that the algorithms presented in the paper have initial convergence speeds that are very close to that of the method in [5]. However, the squared norm of the misalignment vector is more than 12 dB smaller for our algorithm after 20,000 iterations.

In Figure 3, we have plotted the mean behavior of the convergence sequence $\mu(n)$ for the same problem. We can see that $\mu(n)$ goes up very fast initially and then comes down slowly and smoothly. This behavior explains the fast convergence and low misadjustment associated with the algorithms. Also notice the close agreement between the theoretical and empirical results.

Simulation Example #2 In this example, we consider the identification of a time-varying system. The time-varying coefficients of the system are modeled using a random disturbance process as in (5). The initial values of the optimal coefficients were as in example #1 and the disturbance process variance is 10^{-4}. In Figure 4, we have plotted the mean-squared norm of the misalignment vector obtained using the simulations as well as the analysis of Section II. Once again, note that the analysis predicts the behavior of the adaptive filter quite well. The straight-line curve in the figure corresponds to the steady-state value of the squared norm of the misalignment vector when the basic LMS algorithm was used with the optimal choice of the convergence constant μ as given in [9]. Note that our algorithm is able to achieve close to the best performance level possible by the LMS algorithm without really having to pick the optimal value of μ; or really having to know about the statistics of the operating environment.

Simulation Example #3 In this example, we consider identifying a nonlinear system whose input-output relationship is given by a second-order Volterra expansion

$$y(n) = \sum_{i=0}^{3} a_i x(n-i) + \sum_{i=0}^{3}\sum_{j=0}^{i} b_{ij} x(n-i)x(n-j)\ . \quad (31)$$

where

$$\{a_i;\ i = 0, 1, 2, 3\} = \{0.8, 0.6, 0.4, 0.2\} \quad (32)$$

and

$$\begin{bmatrix} b_{0,0} & & & \\ b_{1,0} & b_{1,1} & & \\ b_{2,0} & b_{2,1} & b_{2,2} & \\ b_{3,0} & b_{3,1} & b_{3,2} & b_{3,3} \end{bmatrix} = \begin{bmatrix} 1 & & & \\ .6 & .8 & & \\ .4 & .5 & .6 & \\ .2 & .3 & .4 & .4 \end{bmatrix}. \quad (33)$$

Note that since the output is linear in $\{x(n - i);\ i = 0, 1, 2, 3\}$ and $\{x(n - i)x(n - j);\ i = j, j + 1, ..., 3, j = 0, 1, 2, 3\}$, extension of the basic algorithm to the nonlinear problem is straightforward. Figure 5a displays the mean-squared norm of the misalignment vector corresponding to the linear coefficients and Figure 5b is the corresponding curves for the quadratic coefficients for both the algorithms. We can see that both the algorithms perform very well and about the same in this application.

IV. CONCLUDING REMARKS

This paper presented two stochastic gradient adaptive filtering algorithms with time-varying step sizes. The algorithms are different from traditional methods involving time-varying step sizes in that the changes in the step sizes were also controlled by a gradient algorithm designed to minimize the squared estimation error. We presented a theoretical performance analysis of one of the algorithms in this paper. Experimental results showed that (1) the initial convergence rate of the adaptive filters is very fast. After an initial period when the step size increases very rapidly, the step size decreases slowly and smoothly, given rise to small misadjustment errors; and, (2) in the case of nonstationary environments, the algorithms seek to adjust the step sizes in such a way as to obtain close to the best possible performance. The good properties and the computational simplicity associated with our algorithms makes us believe that they will be used consistently and successfully in several practical applications in the future.

REFERENCES

1. T. J. Shan and T. Kailath, "Adaptive Algorithms with an Automatic Gain Control Feature," *IEEE Trans. Circuits and Systems*, Vol. CAS-35, No. 1, pp. 122-127, January 1988.
2. S. Karni and G. Zeng, "A New Convergence Factor for Adaptive Filters," *IEEE Trans. Circuits and Systems*, Vol. CAS-36, No. 7, pp. 1011-1012, July 1989.
3. C. P. Kwong, "Dual Sign Algorithm for Adaptive Filtering," *IEEE Trans. Communications*, Vol. COM-34, No. 12, pp. 1272-1275, December 1986.
4. R. W. Harris, D. M. Chabries, and F. A. Bishop, *"A Variable Step (VS) Adaptive Filter Algorithm,"* IEEE Trans. Acoustics, Speech, and Signal Proc., Vol. ASSP-34, No. 2, pp. 309-316, April 1986.
5. W. B. Mikhael, *et al.*, "Adaptive Filters with Individual Adaptation of Parameters," *IEEE Trans. Circuits and Systems*, Vol. CAS-33, No. 7, pp. 677-685, July 1986.
6. Y. K. Sin and C. K. Lee, "A Study on the Fast Convergence Algorithm for the LMS Adaptive Filter Design," *Proc. KIEE*, Vol. 22, No. 6, pp. 76-82, November 1985.
7. V. J. Mathews and S. H. Cho, "Improved Convergence Analysis of Stochastic Gradient Adaptive Filters using the Sign Algorithm," *IEEE Trans. Acoustics, Speech, Signal Proc.*, Vol. ASSP-35, No. 4, pp. 450-454, April 1987.
8. A. Feuer and E. Weinstein, "Convergence Analysis of LMS Filters with Uncorrelated Gaussian Data," *IEEE Trans. Acoustics, Speech, Signal Proc.*, Vol. ASSP-33, No. 1, pp. 222-230, February 1985.
9. B. Widrow, *et al.*, "Stationary and Nonstationary Learning Characteristics of the LMS Adaptive Filters," *Proc. IEEE*, Vol. 64, No. 8, pp. 1151-1162, August 1976.

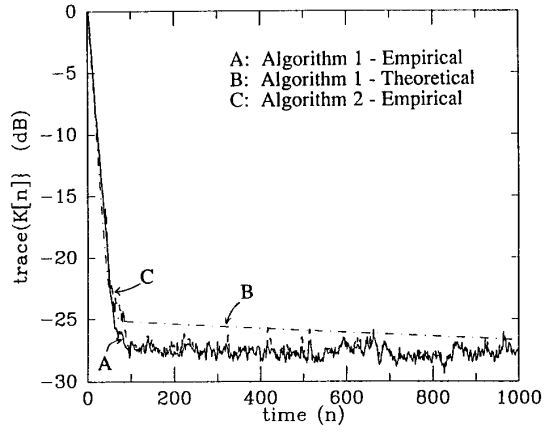


Figure 1. Performance measures for the adaptive filters operating in stationary environments.

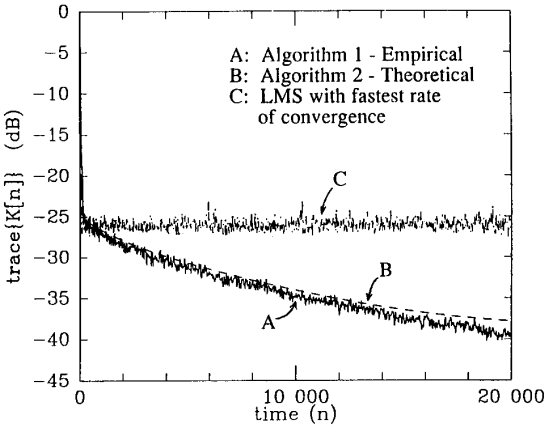


Figure 2. Comparison of the performances of algorithm 1 and the LMS filter designed to give the fastest rate of convergence [5].

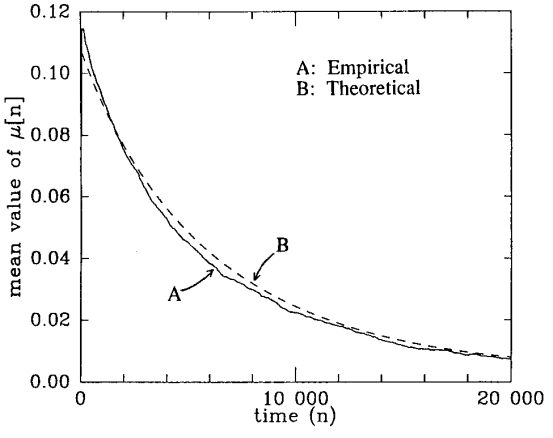


Figure 3. Mean behavior of μ(n) for algorithm 1.

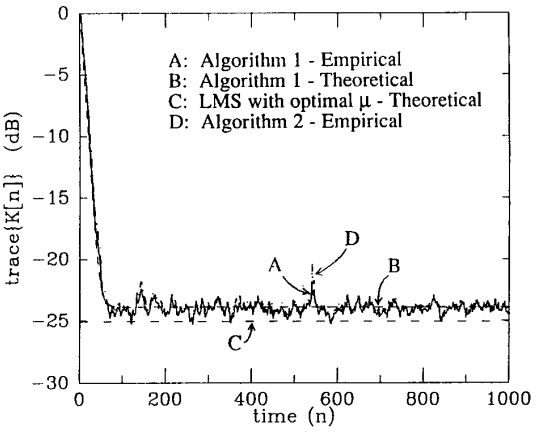
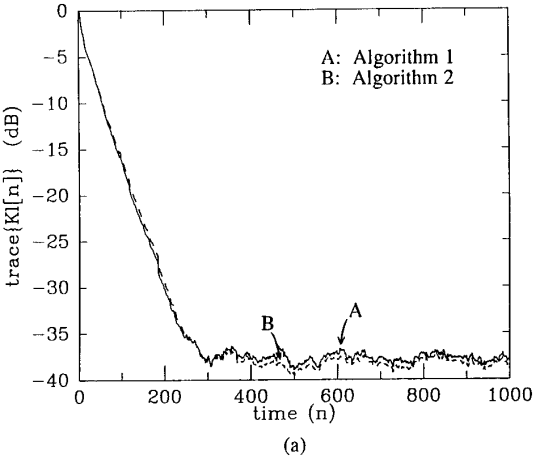


Figure 4. Performance measures for the adaptive filters operating in nonstationary environments.



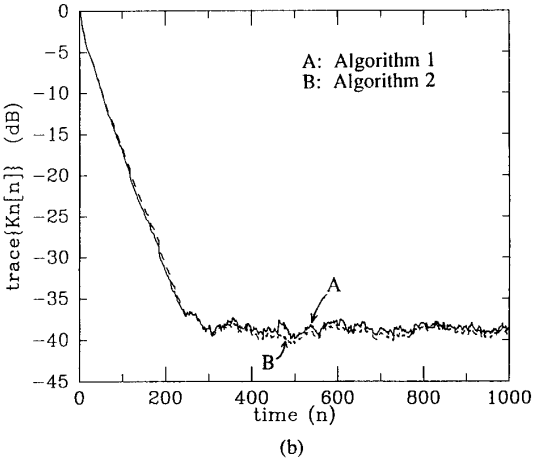


Figure 5. Performance measures of the adaptive filters in nonlinear filtering applications. (a) Squared norm for linear coefficient misalignment vector and (b) squared norm for quadratic coefficient misalignment vector.