EXPONENTIALLY CONVERGENT BEHAVIOUR OF SIMPLE STOCHASTIC ADAPTIVE ESTIMATION ALGORITHMS†

Robert R. Bitmead
Department of Electrical Engineering
University of Newcastle
New South Wales, 2308
Australia

Brian D.O. Anderson
Department of Electrical Engineering
University of Newcastle
New South Wales, 2308
Australia

Abstract

A stochastic algorithm, familiar from adaptive estimation, is introduced and its homogeneous part is shown to be exponentially convergent for a wide class of inputs, which need not be stationary. The implications of this convergence rate for the non-homogeneous algorithm in practical situations are qualitatively examined and a possible approach to improving performance in use is suggested.

1. Introduction

In problems of parameter estimation for stochastic dynamic systems, iterative algorithms have found great application and, indeed, are at the heart of any recursive (on-line) procedure for the adaption of initial estimates to 'true' parameter values and the tracking of time-varying parameters. Their effectiveness in a broad variety of practical situations is well known and the volume of statistical and engineering literature regarding their convergence is substantial.

We consider here a variant of the familiar stochastic approximation scheme. It is well known that this scheme converges with probability one and in mean square for statistically stationary processes provided the convergence parameters tend to zero at specified rates. For the case of slowly time-varying systems it is often desirable to have these parameters tend to a small constant, say $\mu > 0$, in order that changes with time may be tracked. We concentrate on the performance of algorithms with a constant μ.

As a background for our theoretical study we introduce the following filtering problem.

A time series $\{x_k\}$ and a time series $\{y_k\}$ are known to be related in some way, and we seek to model their relation via a moving-average equation of the type

$$\begin{aligned} y_k &= w_0 x_k + w_1 x_{k-1} + \cdots + w_{N-1} x_{k-N+1} \\ &= (x_k \; x_{k-1} \; \cdots \; x_{k-N+1})(w_0 w_1 \cdots w_{N-1})' \\ &\triangleq \underset{\sim}{x}_k' \underset{\sim}{w}^* \end{aligned} \tag{1}$$

The value of $\underset{\sim}{w}^*$ is unknown, and in fact the real relation between $\{x_k\}$ and $\{y_k\}$ may not be of this form. A value of $\underset{\sim}{w}^*$ is sought; just what the significance of $\underset{\sim}{w}^*$ might be in case the real situation cannot be modelled as above is a question we leave for the moment in abeyance, except to comment that (1) would be in some way an approximation, or even best approximation to the true situation.

Now suppose that at time k an estimate $\underset{\sim}{w}_k$ of $\underset{\sim}{w}^*$ is available. Defining an error sequence e_k by

$$e_k = y_k - \underset{\sim}{x}_k' \underset{\sim}{w}_k$$

[which is $\underset{\sim}{x}_k'(\underset{\sim}{w}^* - \underset{\sim}{w}_k)$ in case (1) is exact] we adjust $\underset{\sim}{w}_k$ by some function of e_k to obtain an updated estimate $\underset{\sim}{w}_{k+1}$. The adjustment procedure is so designed that in case (1) is an exact model of reality, $\underset{\sim}{w}_k \to \underset{\sim}{w}^*$, and the convergence properties of the algorithm are fairly easily established. We are however interested in how these algorithms will behave in case (1) is not an accurate description of reality, because for example N is chosen too small, or $\underset{\sim}{w}^*$ is (slowly) time-varying.

The behaviour of the algorithms has been considered by many authors, theoretically and experimentally. For example, for the LMS algorithm

$$\underset{\sim}{w}_{k+1} = \underset{\sim}{w}_k + \mu e_k \underset{\sim}{x}_k \tag{2}$$

there have been several results proved regarding convergence when N is too small. Widrow et al [1] prove many useful results on the convergence of $\underset{\sim}{w}_k$ to the Wiener solution in the case that $\{\underset{\sim}{x}_k\}$ is an independently and identically distributed process* and, although this is a very restrictive assumption on $\{\underset{\sim}{x}_k\}$, they quantify several rules of thumb.

---

*Since $\underset{\sim}{x}_k = [x_k x_{k-1} \cdots x_{k-N+1}]'$, it is apparently impossible when $N > 1$ to secure independence of $\underset{\sim}{x}_k$, $\underset{\sim}{x}_\ell$ for all $k \neq \ell$. However, work of [2] shows how independence of $\{\underset{\sim}{x}_k\}$ can be relaxed to independence of $\{x_k\}$.

---

†Work supported by Australian Research Grants Committee.

Kim and Davisson [3] have proved that for a stationary M-dependent process† $\{x_k\}$ the mean squared error between w_k and the Wiener solution can be made arbitrarily close to zero by choosing μ small enough. And finally Daniell [4] has proved similar results to [3] under assumptions of ergodicity of second moments, boundedness of conditional fourth moments, and asymptotic independence of $\{x_k\}$. In [3,4] the performance of the LMS scheme under time varying conditions is only inferred from convergence, although the analysis of [1] presents a detailed investigation of nonstationary behaviour under the assumption of independent $\{x_k\}$. This investigation shows the LMS algorithm to be exponentially convergent in the mean and the subsequent results follow largely from this.

We follow here a similar pattern: we present the particular algorithm

$$\underset{\sim}{w}_{k+1} = \underset{\sim}{w}_k + \mu e_k \frac{\underset{\sim}{x}_k}{\|\underset{\sim}{x}_k\|^2} \tag{3}$$

and demonstrate in Section II that, for certain broad classes of $\{x_k\}$ which encompass dependent sequences, (3) is exponentially convergent to $\underset{\sim}{w}^*$ with probability one in the case that $\{y_k\}$ is actually a moving average of order N of $\{x_k\}$. Then in Section III we investigate the consequences of exponential convergence of the homogeneous part of (3) in the situations of time variation and of too small an N. Here we utilise the bounded input/bounded output property of an exponentially stable system with a driving term. Section IV contains conclusions and directions for future research.

The study of deterministic adaption algorithms with exponential rates of convergence has been considered by Morgan and Narendra, by Kriesselmeier and by Anderson in [5], [6] and [7,8] and the necessary and sufficient conditions for this rate can be seen to correspond closely to the sufficient conditions that are derived in Section II for the exponential convergence of the stochastic algorithm. Both latter authors have used the deterministic algorithms to devise exponentially convergent adaptive identifiers and observers.

By exponential convergence of a random variable $\{z_k\}$ to zero we mean that the related random variable $\{(1+\beta)^k z_k\}$ for some $\beta > 0$ converges to zero as $k \to \infty$. In this case we say that z_k converges exponentially fast with exponent less than or equal to $-\ln(1+\beta)$. Kushner [9] has a definition of exponential convergence of a random variable which is equivalent to demanding that $\{(1+\beta^k) z_k\}$ be a positive supermartingale.

Nagumo and Noda [2] examine (3) for the homogeneous case where $y_k = \underset{\sim}{x}_k' \underset{\sim}{w}^*$,

$$\underset{\sim}{w}_{k+1} - \underset{\sim}{w}^* = \underset{\sim}{w}_k - \underset{\sim}{w}^* - \mu \frac{\underset{\sim}{x}_k \underset{\sim}{x}_k'}{\|\underset{\sim}{x}\|^2} (\underset{\sim}{w}_k - \underset{\sim}{w}^*) \tag{4}$$

or, writing $\underset{\sim}{v}_k = \underset{\sim}{w}_k - \underset{\sim}{w}^*$,

$$\underset{\sim}{v}_{k+1} = \left(I - \mu \frac{\underset{\sim}{x}_k \underset{\sim}{x}_k'}{\|\underset{\sim}{x}_k\|^2}\right) \underset{\sim}{v}_k \tag{5}$$

They present general results on the convergence of (5) with probability one. They also prove exponential convergence when $\{x_k\}$ is i.i.d. but do not proceed past this stage.

Polyak [10,11] examines the convergence and convergence rate of stochastic algorithms from a general viewpoint and is able to prove exponential rates although his first assumption on the input process represents a benign form of independence. Again, he establishes convergence rates and convergence in various probabilistic senses but does not consider their implications in these papers.

Finally, Sondhi and Mitra [12] have demonstrated exponential bounds on the performance of a continuous-time version of (5) subject to a mixing condition on the input (this condition is very similar to the deterministic conditions of [5,8]). They present an excellent analysis of the effect of exponential convergence of (5) upon the behaviour of (3) in real situations. The main drawback of their analysis, however, is that their mixing condition is extremely severe and, contrary to their claims, one cannot guarantee that it is satisfied by almost all sample paths of a stationary and ergodic input process with positive definite covariance matrix. We present a much less restrictive condition for exponential convergence.

II. Main Result

In this section we examine the convergence rate of the algorithm (5) of Section I. We demonstrate exponential convergence of (5) under an ergodicity assumption - both upper and lower bounds on the exponent are given. Then the pivotal steps of the proof are examined and an extension of the result is derived where the ergodicity requirement is relaxed.

<u>Theorem 1</u>: Let there be given an ergodic sequence $\{x_k\}$ and let the parameter N-vector sequence $\{\underset{\sim}{v}_k\}$ be derived from the algorithm

$$\underset{\sim}{v}_{k+1} = \begin{cases} \underset{\sim}{v}_k & \text{if } \underset{\sim}{x}_k = 0 \text{ - an event of zero probability} \\ \left(I - \mu \dfrac{\underset{\sim}{x}_k \underset{\sim}{x}_k'}{\underset{\sim}{x}_k' \underset{\sim}{x}_k}\right) \underset{\sim}{v}_k & \text{if } \underset{\sim}{x}_k \neq 0 \end{cases}$$

with initial value $\underset{\sim}{v}_0 (|\underset{\sim}{v}_0| < \infty)$ and $\mu \in (0,2)$. Here, $\underset{\sim}{x}_k = [x_k \; x_{k-1} \; . \; . \; . \; x_{k-N+1}]'$. Then if $E[\underset{\sim}{x}_k \underset{\sim}{x}_k']$ is positive definite

$$|\underset{\sim}{v}_0| \underline{\lambda}^k \leq |\underset{\sim}{v}_k| \leq |\underset{\sim}{v}_0| \overline{\lambda}^k$$

---

†M-dependence is defined as follows. Let A and B be two sets of integers with $\min A - \max B > M$. Then $\{x_\alpha | \alpha \in A\}$ and $\{x_\beta | \beta \in B\}$ are independent.

with probability 1 as $k \to \infty$ for $1 > \overline{\lambda} \geq \underline{\lambda} > 0$

Here $\underline{\lambda} = |1-\mu|$ and

$$\overline{\lambda} = \min_{n > (\mu N)^{-1}} (1-\overline{\beta}_n)^{\frac{1}{n}} \quad \text{where}$$

$$\overline{\beta}_n = \frac{1-(1-\mu)^2}{(1+n\mu)^2} \; E\left[\lambda_{min} \sum_{j=0}^{n-1} \frac{\underset{\sim}{x}_j \underset{\sim}{x}_j'}{\underset{\sim}{x}_j' \underset{\sim}{x}_j}\right]$$

and $\lambda_{min}(.)$ is the minimum eigenvalue.

Proof: See Appendix. The minimization carried out for $\overline{\lambda}$ is not necessary to prove exponential convergence but simply allows us to find the least $\overline{\lambda}$. Actually, if $\{x_k\}$ is ergodic and full rank, then there exists a number p such that $0 < \beta_n < 1$ for all $n > p$ and any of these n will do.

Nagumo and Noda [2] have demonstrated an exponential convergence rate for this algorithm in the case that $\{x_k\}$ is a zero mean independent, identically distributed process. Clearly, in this case we may take $n = N$ for μ big enough in order that $\overline{\beta}_n$ be strictly greater than zero. Also, if the algorithm is used with an M-dependent process $\{x_k\}$ (see [3]) then exponential convergence is again easily demonstrated by considering the expectation over an interval of $n \geq M + N + 1$, again for sufficiently large μ. The principal tactic of the proofs of [2], [3] and here, to demonstrate convergence, is to avoid the explicit introduction of conditional expectations and to then devise representations that rely upon ordinary expectations. This clearly makes the results more applicable because they only require knowledge of certain moments of the $\{x_k\}$ whereas conditional expectations, being random variables themselves, need to have a distribution supposed on the $\{x_k\}$ before any thoughts of analysis can be entertained. Perhaps the major drawback of the work by Daniell [4] is his introduced bounds on conditional fourth moments which often can not be assumed a priori.

We believe that this is one of the most useful general convergence results on this algorithm to date because it makes no assumptions on conditional expectations or on absolute coherence lengths of the input sequence. However, while ergodicity is a handy supposition for the proof of convergence rates it is clearly not necessary, and we are able to produce extensions of theorem 1 by replacing the ergodicity requirement with information about the coherence of the input.

The two pivotal steps of the proof of theorem 1 are, firstly, that the matrix $\frac{\underset{\sim}{x}_j \underset{\sim}{x}_j'}{\underset{\sim}{x}_j' \underset{\sim}{x}_j}$ has one eigenvalue of 1 and N-1 of zero - this is used in line (16) to bound $\underset{\sim}{u}_j' \frac{\underset{\sim}{x}_j \underset{\sim}{x}_j'}{\underset{\sim}{x}_j' \underset{\sim}{x}_j} \underset{\sim}{u}_j$ by $\underset{\sim}{u}_j' \underset{\sim}{u}_j$ - and, secondly, that under the ergodicity assumption the random variables $\frac{1}{j} \sum_{m=0}^{j} \beta_{km}$ converge to the expected value with probability one as $j \to \infty$.

The following result of Cramér and Leadbetter [13] (adapted for discrete-time systems) is of interest in that it gives us an alternate condition for a random variable to have a "time-varying ergodicity in the mean", in the sense that time averages tend to ensemble averages. More precisely, we have:

Lemma 1: [13, p.94]

Let $\{z_k\}$ be a sequence of random variables with means $\overline{z}_k$ and with covariance satisfying

$$|E\{(z_k-\overline{z}_k)(z_{k+\tau}-\overline{z}_{k+\tau}\}| \leq K\frac{k^{\alpha}+(k+\tau)^{\alpha}}{1+\tau^{\beta}}, \; 0\leq 2\alpha<\beta<1 \quad (6)$$

Then $\frac{1}{m}\sum_{i=1}^{m} z_i \to \frac{1}{m}\sum_{i=1}^{m} \overline{z}_i$ with probability one and in mean square as $m \to \infty$. This lemma has an obvious application to our current investigation.

Corollary 1: If the $\{x_k\}$ process is loosely correlated (at least up to fourth moments) so that the $\{\beta_{kj}\}$ sequence satisfies (6), then the algorithm (5) will converge exponentially fast with probability one and in mean square.

This corollary represents a crucial observation because it ensures that with certain nonstationary or nonergodic inputs the validity of our proof of exponential convergence of the homogeneous part of (3) remains.

It should be noted here that the ergodicity assumption of theorem 1 and the loose correlation assumption of corollary 1, together with the assumption of full rank covariances, will imply that always there exists a finite k which has $\overline{\beta}_k$ bounded away from zero. That is, the ergodicity assumption performs a dual role in the proof of exponential convergence.

Finally, we remark that, as the boundedness of eigenvalues of the updating matrix appears to be a critical observation for the proof of exponential convergence, it might be expected that the results above could be applicable to a wider class of algorithm than (3) and, indeed, that exponential convergence may hold for any linear decrescent homogeneous algorithm such as the Inverse State scheme of Kumar, Moore and Evans [14]. This hypothesis has not yet been proved, but the similarity between experimental results of different linear algorithms would suggest some underlying unifying structure.

III. Performance Implications of Exponential Convergence of the Homogeneous Algorithm

As has been noted in [1] and [12], the exponential convergence of the homogeneous algorithm (5) allows us to quantify the behaviour of the algorithm (3). There are three main perturbations to (5) which occur in its use in (3). We assume their independence and treat each singly. This is usually justified.

(a) Time Variation in the 'true' Parameter vector $\underset{\sim}{w}^*$. We suppose here that $\underset{\sim}{w}^*$ takes on different values at different times - denote these $\{\underset{\sim}{w}_k^*\}$. Then we may write (5) as

$$\underset{\sim}{v}_{k+1} = [I - \mu \frac{\underset{\sim}{x}_k \underset{\sim}{x}_k'}{\underset{\sim}{x}_k' \underset{\sim}{x}_k}] \underset{\sim}{v}_k - [\underset{\sim}{w}_{k+1}^* - \underset{\sim}{w}_k^*] \qquad (7)$$

where now $\underset{\sim}{v}_j = \underset{\sim}{w}_j - \underset{\sim}{w}_j^*$. This equation will have a bounded solution for $\|\underset{\sim}{v}_k\|$ as $k \to \infty$, provided $\|\underset{\sim}{w}_{k-1}^* - \underset{\sim}{w}_k^*\| \le w_\Delta^*$ for some fixed w_Δ^* and the conditions for exponential convergence are satisfied. Suppose that $\|\underset{\sim}{v}_k\|$ converges to zero in (5) faster than $(1-\beta)^k$. Then as $k \to \infty$ in (7)

$$\|\underset{\sim}{v}_k\| \le \sum_{i=1}^{k} \|\underset{\sim}{w}_{i-1}^* - \underset{\sim}{w}_i^*\|(1-\beta)^{k-i} \|\underset{\sim}{v}_o\| \doteq \frac{w_\Delta^*}{\beta} \|\underset{\sim}{v}_o\|$$

As the choice of possible β is affected by the value of μ, it is clear that the misalignment between $\underset{\sim}{w}_k$ and $\underset{\sim}{w}_k^*$ as $k \to \infty$ is μ-dependent. One may see from the expression for β from theorem 1, as μ increases from zero, β increases to a certain maximum value and then decreases again. Subject to the constraint $\mu \in (0,1]$, it is straightforward to maximize β, in which case the bound on the error due to time-variation in $\underset{\sim}{w}_k^*$ is minimized.

If w^* suffers a jump change in some or all of its elements, at one time only, then we have exponentially fast reconvergence to the new value. This property is of great importance for situations such as fault detection.

The requirement of bounded variations in w^* is hardly restrictive as the types of variation most commonly met with in practice are (i) step changes, e.g. due to component failure or the detection of a new signal source etc., (ii) slow periodic changes e.g. due to diurnal changes in environment such as with communication channels, (iii) small random variations. These variations may include noise in the parameters, which need not be zero mean, i.e. slow drift may also occur. In situations (i) and (ii) above, particularly for the fault detection problem, it would be desirable to have rapid convergence, while for (iii) this need not be so.

(b) Measurement Noise in $\{y_k\}$.

Suppose that the sequence $\{y_k\}$ is not measurable but the sequence $\{z_k\}$ where $z_k = y_k + n_k$ is measurable. Here $\{n_k\}$ is some noise sequence assumed zero mean, white and independent from $\{x_k\}$. Then (3) becomes

$$\underset{\sim}{v}_{k+1} = (I - \mu \frac{\underset{\sim}{x}_k \underset{\sim}{x}_k'}{\underset{\sim}{x}_k' \underset{\sim}{x}_k}) \underset{\sim}{v}_k + \mu \frac{\underset{\sim}{x}_k}{\underset{\sim}{x}_k' \underset{\sim}{x}_k} n_k \qquad (8)$$

and under the assumption of exponential convergence in mean square of (5), which occurs provided the conditions of corollary 1 or theorem 1 are satisfied, it appears that a relation similar to the following must hold:

$$E\|\underset{\sim}{v}_{k+1}\|^2 \le K(1-\beta)^{2k} E\|\underset{\sim}{v}_o\|^2 + \sum_{j=1}^{k} \mu K(1-\beta)^{2(k-j)} E\left[(\underset{\sim}{x}_j' \underset{\sim}{x}_j)^{-1}\right] E\left[n_j^2\right] \qquad (9)$$

So from (8) and (9) we can see that if n_k has bounded variance & $E[(\underset{\sim}{x}_j'\underset{\sim}{x}_j)^{-1}]$ is bounded then $\underset{\sim}{v}_{k+1}$ has bounded variance. The bound on this variance is clearly μ-dependent (and β-dependent). Moreover, the dependence is different to that applying to the analysis tied to (7).

(c) Truncation Errors.

In fitting an N^{th}- order moving average to $\{y_k\}$ we may be neglecting the effect on y_k of values of $\{x_k\}$ before x_{k-N+1}. That is, we may really have $y_k = \underset{\sim}{w}'\underset{\sim}{x}_k$ for $\underset{\sim}{w}$ and $\underset{\sim}{x}_k$ of dimension $M > N$ (M not necessarily finite). Write this as $y_k = \underset{\sim}{w}^{1\prime}\underset{\sim}{x}_k^1 + \underset{\sim}{w}^{2\prime}\underset{\sim}{x}_k^2$ where $\underset{\sim}{w}^1$ has dimension N and $\underset{\sim}{x}^1 = (x_k, x_{k-1}, \dots, x_{k-N+1})'$ and $\underset{\sim}{w}^{2\prime}\underset{\sim}{x}_k^2$ represents the tail effects of $\underset{\sim}{w}$ and $\underset{\sim}{x}_k^2 = (x_{k-N}, x_{k-N-1}, \dots)$ on y_k.

Then (3) becomes

$$\underset{\sim}{v}_{k+1} = \left[I - \mu \frac{\underset{\sim}{x}_k^1 \underset{\sim}{x}_k^{1\prime}}{\underset{\sim}{x}_k^{1\prime} \underset{\sim}{x}_k^1}\right] \underset{\sim}{v}_k + \mu \frac{\underset{\sim}{x}_k^1}{\underset{\sim}{x}_k^{1\prime} \underset{\sim}{x}_k^1} \underset{\sim}{x}_k^{2\prime} \underset{\sim}{w}^2 \qquad (10)$$

The additive term in (10) will clearly produce a misadjustment in $\underset{\sim}{v}_{k+1}$ from $\underset{\sim}{0}$ or 'bias'. This bias is dependent upon $\mu, \underset{\sim}{w}$ and the statistics of the $\{x_k\}$ and a thorough analysis will be presented elsewhere. However, with knowledge of the statistics of $\{x_k\}$ and of the proportion of energy in the tail of $\underset{\sim}{w}$ a bound should be derivable and the dependence of this bound on μ determined. The bias here is of a different character to that arising in (a) and (b). It may even be helpful, in that it yields an N-vector $\underset{\sim}{w}^*$ which defines a more accurate approximation of $[\underset{\sim}{w}^{1\prime}\underset{\sim}{w}^{2\prime}]'$ than does $\underset{\sim}{w}^1$.

The point to be made from (a), (b) and (c) above is that, in practice, the choice of the adaption constant μ is critical to the ultimate performance of the scheme. This has long been realized and the approach we are heading towards is one in which the choice of μ is set up as an optimization problem where the cost function takes into account tracking ability, error variance and bias.

This optimization approach to identification schemes using stochastic-approximation-like algorithms seems to us better motivated and suited to usage than the classic formulas of Robbins and Munro [15] and Kiefer and Wolfowitz [16] which sacrifice performance over finite intervals for strong convergence with zero error in the limit.

IV. Conclusion

We have presented a familiar algorithm from adaptive estimation and established that, provided the input sequence is ergodic or loosely correlated, the homogeneous part of the algorithm converges to zero exponentially fast. The convergence rate depends on the expected value of the minimum eigenvalue of a sum of random matrices. For full rank input sequences the convergence rate may be bounded away from being arbitrarily slow.

We then gave a mainly qualitative examination of the practical implications of exponential convergence when driving terms were added to the homogeneous algorithm. If these terms are bounded then the bounded input/bounded output property of an exponentially stable system is used to infer boundedness of the error. Similarly, subject to independence assumptions, bounded variance inputs produce bounded variance errors. The performance of the algorithm was shown to be dependent upon μ, the adaption constant, as are the effects of time variation of the parameter, measurement noise, and error involved in truncation of the parameter vector. These effects are not all the same, and the fact they are different offers scope for optimally selecting μ.

The results of section III only start to quantify some heuristic notions about the performance of the adaptive algorithm; there is clearly still a need to develop general rules of thumb, applicable in a wide variety of situations, that will aid in choosing good values of μ in practice. Among these rules could be a simple method of associating performance with μ and the spectrum of $\{x_k\}$ if the input were stationary.

Before these rules could be devised, it would be necessary to provide a more thorough quantitative treatment of the propagation of errors and particularly of the dependence of bias upon μ.

Apart from extending the results as above we believe that the general method of establishing exponential convergence as in theorem 1 should be applicable to prove exponential convergence rates for other algorithms arising in identification.

V. Appendix - Proof of Theorem 1

(i) Upper Exponential Bound

We carry through the proof assuming $\mu\epsilon(0,1]$ rather than $\mu\epsilon(0,2)$ as it allows the notation to be more concise. The extension to $\mu\epsilon(0,2)$ is simple and requires just a few modulus signs to be added at various points.

From (5) we have the following results:

Let

$$\underset{\sim}{u}_k = \underset{\sim}{v}_k - \underset{\sim}{v}_0 = (\underset{\sim}{v}_k - \underset{\sim}{v}_{k-1}) + (\underset{\sim}{v}_{k-1} - \underset{\sim}{v}_{k-2}) + \ldots + (\underset{\sim}{v}_1 - \underset{\sim}{v}_0)$$

$$= -\sum_{j=0}^{k-1} \mu \frac{\underset{\sim}{x}_j \underset{\sim}{x}_j'}{\underset{\sim}{x}_j' \underset{\sim}{x}_j} \underset{\sim}{v}_j \tag{11}$$

By an extended application of Cauchy's inequality we have $(\sum_{i=0}^{k-1} \underset{\sim}{a}_i)'(\sum_{i=0}^{k-1} \underset{\sim}{a}_i) \le k \sum_{i=0}^{k-1} \underset{\sim}{a}_i' \underset{\sim}{a}_i$ which yields

$$\underset{\sim}{u}_k' \underset{\sim}{u}_k \le k\mu^2 \sum_{j=0}^{k-1} \frac{(\underset{\sim}{x}_j' \underset{\sim}{v}_j)^2}{\underset{\sim}{x}_j' \underset{\sim}{x}_j} \tag{12}$$

upon application to the inner product of (11) with itself.

The evolution equation for $\underset{\sim}{v}_k$ also yields

$$\underset{\sim}{v}_{j-1}' \underset{\sim}{v}_{j-1} - \underset{\sim}{v}_j' \underset{\sim}{v}_j = [1-(1-\mu)^2] \frac{(\underset{\sim}{v}_{j-1}' \underset{\sim}{x}_{j-1})^2}{\underset{\sim}{x}_j' \underset{\sim}{x}_j}$$

and this shows that $\underset{\sim}{v}_j' \underset{\sim}{v}_j$ is decrescent, a fact used below. Also

$$\underset{\sim}{v}_0' \underset{\sim}{v}_0 - \underset{\sim}{v}_k' \underset{\sim}{v}_k = [1-(1-\mu)^2] \sum_{j=0}^{k-1} \frac{(\underset{\sim}{v}_j' \underset{\sim}{x}_j)^2}{\underset{\sim}{x}_j' \underset{\sim}{x}_j} \tag{13}$$

and (12) implies

$$\underset{\sim}{v}_0' \underset{\sim}{v}_0 - \underset{\sim}{v}_k' \underset{\sim}{v}_k \ge [1-(1-\mu)^2] \frac{\underset{\sim}{u}_k' \underset{\sim}{u}_k}{k\mu^2} \tag{14}$$

Equation (14) will be used further below. Meanwhile, rewrite (13) using $\underset{\sim}{u}_k = \underset{\sim}{v}_k - \underset{\sim}{v}_0$.

$$\underset{\sim}{v}_0' \underset{\sim}{v}_0 - \underset{\sim}{v}_k' \underset{\sim}{v}_k = [1-(1-\mu)^2] \sum_{j=0}^{k-1} \frac{[(\underset{\sim}{u}_j + \underset{\sim}{v}_0)' \underset{\sim}{x}_j]^2}{\underset{\sim}{x}_j' \underset{\sim}{x}_j}$$

so that, using $\|\underset{\sim}{a}+\underset{\sim}{b}\| \ge \|\underset{\sim}{a}\| - \|\underset{\sim}{b}\|$ with $(\underset{\sim}{a})_j = \underset{\sim}{u}_j' \underset{\sim}{x}_j$ and $(\underset{\sim}{b})_j = \underset{\sim}{v}_0' \underset{\sim}{x}_j$,

$$(\underset{\sim}{v}_0' \underset{\sim}{v}_0 - \underset{\sim}{v}_k' \underset{\sim}{v}_k)^{\frac{1}{2}} \ge [1-(1-\mu)^2]^{\frac{1}{2}} \left\{ -\left[\sum_{j=0}^{k-1} \frac{(\underset{\sim}{u}_j' \underset{\sim}{x}_j)^2}{\underset{\sim}{x}_j' \underset{\sim}{x}_j} \right]^{\frac{1}{2}} + \left[\sum_{j=0}^{k-1} \frac{(\underset{\sim}{v}_0' \underset{\sim}{x}_j)^2}{\underset{\sim}{x}_j' \underset{\sim}{x}_j} \right]^{\frac{1}{2}} \right\} \tag{15}$$

We next find a bound on the first sum of (15)

$$\sum_{j=0}^{k-1} \frac{(\underset{\sim}{u}_j' \underset{\sim}{x}_j)^2}{\underset{\sim}{x}_j' \underset{\sim}{x}_j} \le \sum_{j=0}^{k-1} \underset{\sim}{u}_j' \underset{\sim}{u}_j \tag{16}$$

$$\le \sum_{j=0}^{k-1} \frac{\underset{\sim}{v}_0' \underset{\sim}{v}_0 - \underset{\sim}{v}_j' \underset{\sim}{v}_j}{[1-(1-\mu)^2]} j\mu^2, \text{ using (14)}$$

$$\le \frac{(\underset{\sim}{v}_0' \underset{\sim}{v}_0 - \underset{\sim}{v}_k' \underset{\sim}{v}_k) k^2 \mu^2}{[1-(1-\mu)^2]}, \text{ because } \underset{\sim}{v}_j' \underset{\sim}{v}_j \text{ is decrescent.}$$

Now (15) yields

$$(\underset{\sim}{v}_0' \underset{\sim}{v}_0 - \underset{\sim}{v}_k' \underset{\sim}{v}_k)^{\frac{1}{2}} \ge -(\underset{\sim}{v}_0' \underset{\sim}{v}_0 - \underset{\sim}{v}_k' \underset{\sim}{v}_k)^{\frac{1}{2}} k\mu + [1-(1-\mu)^2]^{\frac{1}{2}} \left[\underset{\sim}{v}_0' \sum_{j=0}^{k-1} \frac{\underset{\sim}{x}_j \underset{\sim}{x}_j'}{\underset{\sim}{x}_j' \underset{\sim}{x}_j} \underset{\sim}{v}_0 \right]^{\frac{1}{2}}$$

$$\ge -(\underset{\sim}{v}_0' \underset{\sim}{v}_0 - \underset{\sim}{v}_k' \underset{\sim}{v}_k)^{\frac{1}{2}} k\mu + [1-(1-\mu)^2]^{\frac{1}{2}} (\underset{\sim}{v}_0' \underset{\sim}{v}_0)^{\frac{1}{2}} \left[\lambda_{min} \sum_{j=0}^{k-1} \frac{\underset{\sim}{x}_j \underset{\sim}{x}_j'}{\underset{\sim}{x}_j' \underset{\sim}{x}_j} \right]^{\frac{1}{2}}$$

or

$$\frac{\underset{\sim}{v}_0' \underset{\sim}{v}_0 - \underset{\sim}{v}_k' \underset{\sim}{v}_k}{\underset{\sim}{v}_0' \underset{\sim}{v}_0} \ge \left[\frac{1-(1-\mu)^2}{(1+k\mu)^2} \right] \lambda_{min} \left[\sum_{j=0}^{k-1} \frac{\underset{\sim}{x}_j \underset{\sim}{x}_j'}{\underset{\sim}{x}_j' \underset{\sim}{x}_j} \right] \tag{17}$$

As $\frac{1-(1-\mu)^2}{(1+k\mu)^2} < \frac{1}{k^2\mu^2}$ and $\lambda_{min}\left(\sum_{j=0}^{k-1} \frac{\underset{\sim}{x}_j \underset{\sim}{x}_j'}{\underset{\sim}{x}_j' \underset{\sim}{x}_j}\right) \le \frac{k}{N}$ (using the facts that $\lambda_{min}(A) \le \frac{\text{trace } A}{\dim A}$, trace $A+B$ = trace A + trace B and trace $\frac{\underset{\sim}{x}_j \underset{\sim}{x}_j'}{\underset{\sim}{x}_j' \underset{\sim}{x}_j} = 1$), provided $k\mu N > 1$, the right hand side of (17), call it β_k say, is less than one and greater than or equal to zero.

The parameters μ and N are of course fixed.

Let us also temporarily fix k so that $k\mu N > 1$. We consider the sequence $\underset{\sim}{v}_0'\underset{\sim}{v}_0, \underset{\sim}{v}_k'\underset{\sim}{v}_k, \underset{\sim}{v}_{2k}'\underset{\sim}{v}_{2k}, \ldots, \underset{\sim}{v}_{mk}'\underset{\sim}{v}_{mk}, \ldots$ for interger m.

We have established above that $\underset{\sim}{v}_k'\underset{\sim}{v}_k \le (1-\beta_k)\underset{\sim}{v}_0'\underset{\sim}{v}_0$ and more generally we have

$$\underset{\sim}{v}_{mk}'\underset{\sim}{v}_{mk} \le (1-\beta_{mk})\underset{\sim}{v}_{(m-1)k}'\underset{\sim}{v}_{(m-1)k}$$

Setting $\tau_{m+1} = (1-\beta_{mk})\tau_m$ and $\tau_0 = \underset{\sim}{v}_0'\underset{\sim}{v}_0$ we observe that $0 \le \underset{\sim}{v}_{mk}'\underset{\sim}{v}_{mk} \le \tau_m$ and prove an exponential convergence to zero of τ_m.

Clearly

$$\tau_m = \left[\prod_{i=0}^{m-1}(1-\beta_{ik})\right]\tau_0 \quad \text{and taking logarithms}$$

$$\ln \tau_m - \ln \tau_0 = \sum_{i=0}^{m-1} \ln(1-\beta_{ik}) \qquad (18)$$

The requirement that τ_m converge to zero regardless of initial conditions is that the right hand side of (18) diverge to minus infinity. Furthermore, if this term diverges faster than $m\ln(1-\alpha)$ for $\alpha\in(0,1)$ then τ_m converges exponentially fast to zero with exponent less than $\ln(1-\alpha)$. These requirements are clearly satisfied provided $\{\beta_{ik}\}$ is bounded away from zero (infinitely) often enough.

We now make the observation that if $\{x_k\}$ is ergodic then so is $\{\beta_{ik}\}$ and also $\{\ln(1-\beta_{ik})\}$. And, if $\{\ln(1-\beta_{ik})\}$ is ergodic, then

$$m^{-1}\sum_{i=0}^{m-1} \ln(1-\beta_{ik}) \to E[\ln(1-\beta_{mk})]$$

with probability one as $m\to\infty$.

Applying Jensen's inequality

$$E[\ln(1-\beta_{mk})] \le \ln(1-E(\beta_{mk})] = \ln(1-\bar{\beta}_k)$$

Further, if $\underset{\sim}{x}_k$ is ergodic and has full rank covariance then $\frac{1}{\ell}\sum_{i=0}^{\ell-1}\frac{\underset{\sim}{x}_i\underset{\sim}{x}_i'}{\underset{\sim}{x}_i'\underset{\sim}{x}_i}$ converges to some full rank matrix as $\ell\to\infty$ so that there always exists some integer p such that, provided $k > p$, $\bar{\beta}_k > 0$.

Thus τ_m converges to zero with probability one and at an exponential rate faster than $(1-\bar{\beta}_k)^m$. Now, although the $\{\tau_m\}$ sequence only captures every k^{th} element of $\underset{\sim}{v}_\ell'\underset{\sim}{v}_\ell$, the fact that $\underset{\sim}{v}_\ell'\underset{\sim}{v}_\ell$ is decrescent shows that $\underset{\sim}{v}_m'\underset{\sim}{v}_m$ converges exponentially fast to zero with rate faster than

$$(1-\bar{\beta}_k)^{\frac{m}{k}}$$

<u>(ii) Lower Exponential Bound</u>

From (5) we have, since the eigenvalue of minimum magnitude of

$$\left(I - \mu\frac{\underset{\sim}{x}_k\underset{\sim}{x}_k'}{\underset{\sim}{x}_k'\underset{\sim}{x}_k}\right) \text{ has magnitude } |1-\mu|, \text{ that}$$

$$\|\underset{\sim}{v}_{k+1}\| \ge |1-\mu|\;\|\underset{\sim}{v}_k\|$$

whence $\|\underset{\sim}{v}_{k+1}\| \ge |1-\mu|^{k+1}\|\underset{\sim}{v}_0\|$. Taking $\underline{\lambda} = |1-\mu|$ we have the final part of the theorem, provided $0 < \mu < 2$.

∇∇∇

REFERENCES

[1] B. Widrow, J. McCool, M.G. Larimore and C.R. Johnson Jr., "Stationary and nonstationary learning characteristics of the LMS adaptive filter", *Proc. IEEE*, vol.64, No.8, August 1976, pp. 1151-1162.

[2] J-I. Nagumo & A. Noda, "A Learning method for system identification", *IEEE Trans.Auto.Control* Vol.AC-12, No.3, June 1967, pp.282-287.

[3] J-K. Kim & L.D. Davisson, "Adaptive Linear Estimation for stationary M-dependent processes", *IEEE Trans. Inf. Theory*, Vol.IT-21,No.1, Jan. 1975, pp. 23-31.

[4] T.P. Daniell, "Adaptive estimation with mutually correlated training sequences", *IEEE Trans. Sy.Sci. & Cyb.*, Vol.SSC-6,No.1, January 1970, pp.12-19.

[5] A.P. Morgan & K.S. Narendra, "On the Uniform Asymptotic Stability of Certain Linear Nonautonomous Differential Equations", *SIAM J.Cntrl.& Optimization*, Vol.15,No.1, Jan.1977,pp.5-24.

[6] G. Kreisselmeier, "Adaptive observers with exponential rate of convergence",*IEEE Trans.Auto Cntrl.*,Vol.AC-22,No.1, Feb.1977, pp.2-8.

[7] B.D.O. Anderson, "Exponential stability of linear equations arising in adaptive identification", *IEEE Trans.Auto.Cntrl*, Vol.AC-22,No.1, February 1977, pp. 83-88.

[8] B.D.O. Anderson, "Adaptive identification of multiple-input,multiple-output plants", *Proc. 1974 IEEE Conf. Dec. & Cntrl*, pp.273-281.

[9] H.J. Kushner, <u>Introduction to Stochastic Control</u>, Holt, Rinehart & Winston Inc, 1971.

[10] B.T. Polyak, "Convergence & convergence rate of iterative stochastic algorithms, I. General Case", *Avt. i Telemekh.*, No.12, December 1976 pp. 1858-1868.

[11] B.T. Polyak, "Convergence & convergence rate of iterative stochastic algorithms, II. The Linear case", *Avt. i Telemekh.*, No.4, April 1977, pp. 537-542.

[12] M.M. Sondhi & D. Mitra, "New results on the performance of a well-known class of adaptive filters", *Proc.IEEE*, Vol.64, No.11, November 1976, pp. 1583-1597.

[13] H. Cramér & M.R. Leadbetter, <u>Stationary and Related Stochastic Processes</u>, John Wiley and Sons, Inc., 1967

[14] R. Kumar, J.B. Moore & R.J. Evans, "Inverse state stochastic approximation recursions", Preprint Elec. Eng. Dept. Univ. of Newcastle, Australia, submitted for publication.

[15] H. Robbins & S. Munro, "A Stochastic approximation method", *Ann.Math.Stat.*, Vol. 22, 1951, pp. 400-407.

[16] J. Kiefer & J. Wolfowitz, "Stochastic Estimation of the maximum of a Regression Function", *Ann.Math.Stat.* Vol.23,1952, pp. 462-466.