Digital Adaptive Filters: Conditions for Convergence, Rates of Convergence, Effects of Noise and Errors Arising from the Implementation

ALAN WEISS AND DEBASIS MITRA, MEMBER, IEEE

***Abstract*—A variety of theoretical results are derived for a well-known class of discrete-time adaptive filters. First the following idealized identification problem is considered: a discrete-time system has vector input $x(t)$ and scalar output $z(t)=h'x(t)$ where h is an unknown time-invariant coefficient vector. The filter considered adjusts an estimate vector $\hat{h}(t)$ in a control loop according to**

$$\hat{h}(t+\Delta t)=\hat{h}(t)+K[z(t)-\hat{z}(t)]x(t),$$

where $\hat{z}(t)=\hat{h}(t)'x(t)$ and K is the control loop gain. The effectiveness of the filter is determined by the convergence properties of the misalignment vector $r(t)=h-\hat{h}(t)$. It is shown that a certain nondegeneracy "mixing" condition on the input $\{x(t)\}$ is necessary and sufficient for the exponential convergence of the misalignment. Qualitatively identical upper and lower bounds are derived for the rate of convergence. Situations where noise is present in $z(t)$ and $x(t)$ and the coefficient vector h is time-varying are analyzed. Nonmixing inputs are also considered, and it is shown that in the idealized model the above stability results apply with only minor modifications. However, nonmixing input in conjunction with certain types of noise lead to bounded input – unbounded output, i.e., instability.

I. INTRODUCTION

IN THIS PAPER we derive a variety of theoretical results for a well-known class of discrete-time adaptive filters. The results obtained here on the conditions for convergence, rates of convergence, and the effects of noise equal in scope results recently obtained for the continuous-time analog counterparts. This paper has the additional purpose of analyzing and elucidating some of the unusual, hitherto unexplained behavior of some advanced realizations in digital hardware that have recently appeared and are in the process of being evaluated.

A. The Adaptation Algorithm

As an introduction to the adaptation algorithm studied here, let us first consider the following idealized identification problem (see Fig. 1). An unknown system (or black box) has a sequence of vector inputs $x(t)$, each of known dimension n, and a sequence of scalar outputs $z(t)$. Both sequences are known, or observed, at times $t=t_0, t_0+\Delta t, t_0+2\Delta t, \cdots$, and it is assumed that

$$z(t)\equiv h'x(t), \tag{1}$$

where h is a constant n-vector and the prime denotes matrix transposition. The problem is to estimate h.

The adaptive procedure starts with an initial estimate $\hat{h}(t_0)$ and recursively adjusts the estimates $\hat{h}(t)$ according to the difference equation

$$\Delta\hat{h}(t) \triangleq \hat{h}(t+\Delta t)-\hat{h}(t)=K\{z(t)-\hat{z}(t)\}x(t), \tag{2}$$

where

$$\hat{z}(t)=\hat{h}(t)'x(t) \tag{3}$$

and K, the control loop gain, is a parameter. It is assumed

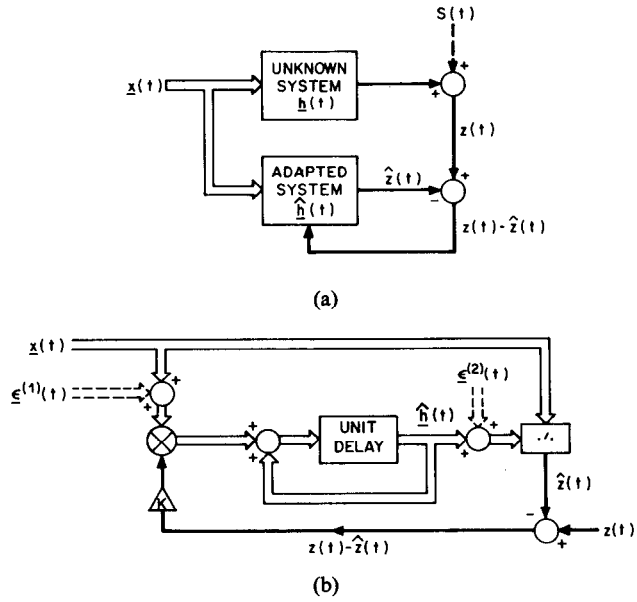

Fig. 1. (a) Schematic of identification problem. In the idealized problem the system noise $s(t)\equiv 0$ and $h(t)$ is constant. In Section IV-B and IV-C these restrictions are removed. (b) Schematic of adaptive filter. The box indicated by $\cdot/\cdot$ refers to the multiplications and a summation involved in forming the scalar product of two vectors. Two kinds of errors arising from the implementation, $\epsilon^{(1)}(t)$ and $\epsilon^{(2)}(t)$, are considered in Section V. In the idealized problem $\epsilon^{(1)}(t)=\epsilon^{(2)}(t)\equiv 0$. The step involved in normalizing the norm of $x(t)$ to unity is not shown.

Manuscript received February 26, 1978; revised March 28, 1979. This paper was presented at the 1978 IEEE International Conference on Acoustics, Speech, and Signal Processing.

A. Weiss was with Bell Laboratories, Murray Hill, NJ. He is now with the Courant Institute of Mathematical Sciences, New York University, Mercer Street, New York, NY.

D. Mitra is with Bell Laboratories, Room 2C-454, Murray Hill, NJ 07974.

0018-9448/79/1100-0637$00.75

throughout that

$$\|x(t)\|^2 \triangleq x'(t)x(t) \equiv 1. \tag{4}$$

Thus a normalization procedure which consists of dividing the right side of (2) by $\|x(t)\|^2$ is tacitly assumed. This normalization procedure is not undertaken in all implementations. Nevertheless we assume (4) both because some implementations (for example, Duttweiler's [20]) do use this normalization, and for mathematical convenience. If instead of (4) we had assumed that

$$\|x(t)\|^2 \leqslant L^2, \tag{4'}$$

as in [1], then our upper bounds in Section II hold with only minor changes. See, for instance, footnote two to inequality (19).

The effectiveness of the filter is determined by the convergence properties of the misalignment vector $r(t)$ which is defined by

$$r(t) \triangleq h - \hat{h}(t). \tag{5}$$

We see that

$$\begin{aligned} \Delta r(t) &= -\Delta\hat{h}(t), \quad \text{since } \Delta h = 0, \\ &= -K\{z(t) - \hat{z}(t)\}x(t), \quad \text{from (2)}, \\ &= -K\{r'(t)x(t)\}x(t). \end{aligned} \tag{6}$$

The convergence properties of the solutions $r(t)$ of the above homogeneous difference equation are the subject of the analysis reported in Sections II and III. Our discussion in Sections I-B and I-C will indicate that because of the robustness and simplicity of the algorithm it has found a variety of applications. However, the results hitherto available leave unresolved some of the basic questions regarding the performance of the algorithm. Some of these questions are "What is the least stringent condition on the input vectors $\{x(t)\}$ which guarantees uniform convergence of the misalignment? What are the rates of convergence when the input belong to the class for which convergence is guaranteed?" These questions are germane when the input vectors are derived from a complex signal such as speech.

B. Summary of Results

In this paper we answer some of these questions. In Section I-E we define the key notion of 'mixing' input. We emphasize that our usage of the term mixing is not to be confused with other usages. This is the discrete-time analog of the mixing condition introduced in [1] and [2] for continuous-time vector processes. In particular, mixing does not require stationarity or periodicity of the input signal, or even that it is either stochastic or deterministic. We are able to prove the following results:

i) We show in Section II that the mixing condition implies the existence of an exponentially decreasing upper bound on $\|r(t)\|$. We also show in Appendix I that the existence of an exponentially decreasing upper bound implies that the mixing condition is satisfied. Thus the mixing condition is necessary and sufficient for exponential convergence of the misalignment.

ii) The upper bound on the rate of convergence is valid for all mixing inputs, all K and all t. The mixing condition is also used to obtain a lower bound on the convergence rate for small values of K. A related but not identical assumption is used in Section III to derive a lower bound for larger values of K. The motivation for the care that is taken to obtain these bounds is that it provides important insights into the question of the best loop gain setting. This is exemplified by the fact that the upper and lower bounds have identical qualitative dependence on K for both small and large K.

iii) We conclude that, for the large class of processes for which both the upper and lower bounds apply, the rate of convergence must increase as K for small K and must decrease as $1/K$ to a very small number as $K \to 1$, since this behavior is common to both bounds. The value of K which maximizes the rate of convergence for our upper bound is a small number, considerably smaller than the "optimum" value of K predicted by a myopic optimization and the "method of averaging".

In Section IV we proceed to investigate the effects of adding a vector forcing term $u(t)$ to the right side of (6), i.e.,

$$\Delta r(t) = -K\{r'(t)x(t)\}x(t) + u(t). \tag{7}$$

iv) We show that if $\|u(t)\|$ is bounded, or equivalently has a bounded mean over intervals of a finite length, then so is $\|r(t)\|$. In particular, the residual error $\|r(t)\|$ is bounded as $t \to \infty$. Explicit bounds for the residual error are obtained so that its dependence on the loop gain is transparent.

v) The above bounded input – bounded output property is exploited by noting that the effects of departures from the idealized problem can be represented by the term $u(t)$. Thus: a) the effect of an added system noise component $s(t)$ in the observed signal $z(t)$, and b) the effect of variations with time of the unknown vector h, can both be lumped into the term $u(t)$ in (7).

vi) In the final part of the paper, Section V, we consider for the first time the rather consequential implications of nonmixing inputs on the performance of the filter. We begin by showing that inputs may be expected to be nonmixing in many applications, especially communications-related applications such as echo-cancellation. In these cases the high dimensionality of the input vectors and the filter, together with the bandlimited form of the inputs, are responsible for the phenomenon of nonmixing inputs. We show that in the idealized problem as well as in the case where noise $s(t)$ is present in the measured signal (case a) above) the results obtained previously on the basis of the mixing assumption on the input vectors apply (with only minor modifications) to the case of nonmixing inputs.

vii) The situation changes abruptly if nonmixing inputs are considered in conjunction with random errors of two different kinds that may arise due to noise or a digital implementation of the device. We prove the surprising and consequential result that if both kinds of errors occur simultaneously, each with arbitrarily small bounds, then

$\|r(t)\|$ becomes unbounded as $t\to\infty$. If, however, only one kind of error occurs, then the residual error is bounded provided the bound on the error and the loop gain is small.

C. Applications

Eykhoff [3] provides an authoritative account of the variety of approaches to the identification problem as well as the applications that the algorithm has found. The algorithm has been proposed for adapting switching circuits [4], control [5], [7], and self-optimization [6]. Among communication related applications is the equalization of telephone lines for data communication [8], [9], [10]. The algorithm has been proposed for echo cancellation in long distance telephony [11]–[14]. Both analog [15], [16] and digital [17]–[20] versions of the canceller have been realized. (A point to note about the cancellers is that typically the dimension n is large being on the order of 100.) Speech related applications are to be found in [21] and [22].

D. Known Theoretical Results

A key equation derived from (6) helps to explain the basic robustness of the algorithm:

$$\Delta\|r(t)\|^2 \equiv \|r(t+\Delta t)\|^2 - \|r(t)\|^2 = -K(2-K)\{r'(t)x(t)\}^2. \tag{8}$$

For

$$\begin{aligned}\Delta\|r(t)\|^2 &= \{r(t+\Delta t)-r(t)\}'\{r(t+\Delta t)+r(t)\}\\ &= \{r(t+\Delta t)-r(t)\}'\{r(t+\Delta t)-r(t)+2r(t)\}\\ &= \|r(t+\Delta t)-r(t)\|^2+2r'(t)\{r(t+\Delta t)-r(t)\}\end{aligned} \tag{9}$$

which yields (8) when the expression for $\Delta r(t)$ in (6) is substituted into the expression on the right side.

Equation (8) says that for $0<K<2$, the norm of the misalignment is nonincreasing.[1] This is of course not the same as uniform convergence of $\|r(t)\|$ to zero; additional information is called for regarding the behavior of the term $r'(t)x(t)$. We note from (6) that choosing $K=1+\delta$ has virtually the same effect as choosing $K=1-\delta$; in either case the norm of the component of $r(t+\Delta t)$ in the $x(t)$ direction will be $|\delta\{r'(t)x(t)\}|$. So henceforth we shall assume that $0<K\leqslant 1$.

Equation (8) is also noteworthy because it focuses on a fundamental difference between continuous- and discrete-time versions of the adaptive filter; in the former case the misalignment norm is nonincreasing for all values of the loop gain K (see for instance [1]). However, we shall find that there is remarkable affinity between the results proved in Sections II, III, and IV, and the corresponding results for continuous-time filters [1] provided an appropriate change in the scale of K is taken into account; thus large K is interpreted as K approaching unity and K approaching infinity for the discrete-time and continuous-time filters, respectively.

Some of our results have been initiated by the methods and results presented in two recent papers on the continuous-time algorithm. Our derivations of the upper bound, and the subsequent results on the solutions of (7) which includes the forcing term $u(t)$, are adaptations of the methods in [1]. In an important paper Morgan and Narendra [2] proved that the mixing condition is not only sufficient but also necessary for exponential convergence in continuous-time. Our proof in the Appendix of the necessity of the mixing condition for uniform convergence is an adaptation of the proof provided in [2]. On the other hand, the derivation of the lower bound for large K, based as it is on geometrical arguments, is basically new. Also, almost all the results in Sondhi and Mitra's paper, including their lower bound, may be derived from the results given here by going to the continuous limit in an appropriate manner. Finally, the results in Section V concern topics which have virtually not been addressed previously in either the continuous- or discrete-time formulations. Thus the implications of nonmixing inputs have not been investigated previously; we find that the implications are rather consequential. We also recall that there is a considerable body of literature concerning the behavior of the algorithm under a variety of assumptions regarding the input vectors [7], [9], [23]–[25].

As far as convergence rates are concerned, all published results are essentially based upon averaging of the right side of (6) and (7) and assuming r to be either slowly varying or independent of x [26], [27], [28]. Some of the early results on the method of averaging were established for the deterministic, continuous time equations by Bogoliubov [29]. Khasminskii [30] has shown that the method of averaging provides uniformly good approximations to the true solutions over intervals of order $1/K$ in the continuous-time formulation. However, the method of averaging gives misleading results in all cases except where K is very small.

E. The Mixing Condition

As mentioned above almost all our results require familiarity with the mixing condition on the input vectors $\{x(t)\}$. The following is the discrete analog of the mixing condition in [1].

The vectors $x(t)$ satisfy the mixing condition if there exist numbers T and $\alpha>0$ such that for any constant nonzero n-vector d and any time t,

$$\frac{1}{T}\sum_{j=0}^{T-1}\{d'x(t+j\Delta t)\}^2 \geqslant \alpha\|d\|^2. \tag{10}$$

An equivalent statement of the mixing condition is the following discrete analog of the condition used in [2].

[1] Without the assumption $\|x(t)\|=1$, we should have obtained $\Delta\|r(t)\|^2 = -K\{2-K\|x(t)\|^2\}\{r'(t)x(t)\}^2$; thus a decreasing misalignment is implied only if $0<K<2/\|x(t)\|^2$.

There exist numbers $a>0$ and b such that for any unit n-vector $\boldsymbol{d}$, any time t and any $N \geqslant 1$

$$\sum_{j=0}^{N-1} \{\boldsymbol{d}'\boldsymbol{x}(t+j\Delta t)\}^2 \geqslant aN+b. \tag{11}$$

Let us examine the condition in (10) in greater detail. This condition is basically that, over any time interval of length T, the components of $\boldsymbol{x}(t)$ have an average length of at least α in any direction. In particular, a sequence $\{\boldsymbol{x}(t)\}$ in which the n-vectors are restricted to any proper subspace of R^n is nonmixing. Further, if the input is nonmixing there will be arbitrarily long time intervals for which the vectors $\boldsymbol{x}(t)$ are effectively restricted to a particular proper subspace of R^n.

It is also clear that where n is the dimension of $\boldsymbol{x}(t)$,

$$T \geqslant n \quad \text{and} \quad \alpha \leqslant 1/n. \tag{12}$$

The first follows from observing that it takes at least n vectors to span an n-dimensional space. For the second inequality observe that the smallest average component of a collection of n-vectors can be no larger than $1/n$. A better proof is to note that there is no loss of generality in assuming that

$$\alpha = \text{smallest eigenvalue of } \frac{1}{T}\sum_{j=0}^{T-1} \boldsymbol{x}(t+j\Delta t)\boldsymbol{x}'(t+j\Delta t). \tag{13}$$

Then note that $\|\boldsymbol{x}(t)\|=1$ for all t, so the trace of $\boldsymbol{x}(t)\boldsymbol{x}'(t)$ is unity for each t, and thus the trace of the matrix in (13) is also unity. As the trace is the sum of the n eigenvalues, the smallest eigenvalue cannot exceed $1/n$.

It should be noted in (10) that any $T_1 \geqslant T$ will suffice in the mixing condition, perhaps with a new α, so we should properly regard $\alpha = \alpha(T)$.

It may be seen that many stochastic processes do not satisfy the mixing condition. However, for many processes of interest, there will be choices of T and α such that the sample paths will be mixing for long periods of time separated by periods when the process is not mixing. In the former periods, our exponential bounds will hold while in the latter periods, by virtue of (8), $\|\boldsymbol{r}(t)\|$ is nonincreasing.

For the sake of brevity and simplicity we will agree that, from now on, in all summations the index will be incremented by Δt. Thus

$$\sum_{j=t_0}^{t_0+(T-1)\Delta t} \boldsymbol{x}(j) = \boldsymbol{x}(t_0)+\boldsymbol{x}(t_0+\Delta t)+\cdots+\boldsymbol{x}\{t_0+(T-1)\Delta t\}.$$

II. Upper Bound

In this section we will derive exponentially decreasing upper bounds on the norm of the misalignment vector $\boldsymbol{r}(t)$ in the idealized problem, (6), where $\boldsymbol{x}(t)$ is mixing. Our results and proofs are close to those in Sondhi and Mitra's paper [1, Section IIB].

The better of our upper bounds for small values of α ($\alpha \leqslant 0.05$, $n \geqslant 20$) is also the simplest to derive and evaluate numerically. Another upper bound, which is an improvement only for large values of α, is stated without proof.

A. Derivation of the Bound

The mixing condition seems to say that $\boldsymbol{r}'(t)\boldsymbol{x}(t)$ cannot be small all the time if $\|\boldsymbol{r}(t)\|$ is large. This is exactly what we need, according to (8), for $\|\boldsymbol{r}(t)\|$ to decrease. Equation (10), the mixing condition, leads us to consider $\sum_{t=t_0}^{t_0+(T-1)\Delta t}\{\boldsymbol{r}'(t_0)\boldsymbol{x}(t)\}^2$. We have

$$\begin{aligned}\alpha T\|\boldsymbol{r}(t_0)\|^2 &\leqslant \sum_{t=t_0}^{t_0+(T-1)\Delta t}\{\boldsymbol{r}'(t_0)\boldsymbol{x}(t)\}^2, \quad \text{from (10)},\\ &= \sum_{t=t_0}^{t_0+(T-1)\Delta t}\{\boldsymbol{x}'(t)\boldsymbol{r}(t)\}^2\\ &\quad + \sum_{t=t_0}^{t_0+(T-1)\Delta t}\left[\boldsymbol{x}'(t)\{\boldsymbol{r}(t_0)-\boldsymbol{r}(t)\}\right]^2\\ &\quad + 2\sum_{t=t_0}^{t_0+(T-1)\Delta t}\{\boldsymbol{x}'(t)\boldsymbol{r}(t)\}\left[\boldsymbol{x}'(t)\{\boldsymbol{r}(t_0)-\boldsymbol{r}(t)\}\right].\end{aligned} \tag{14}$$

We now bound each term on the right side of (14) from above, and derive an inequality involving $\|\boldsymbol{r}(t_0)\|$ and $\|\boldsymbol{r}(t_0+T\Delta t)\|$. We will need the following formula, valid for any n-vectors $\boldsymbol{a}(t)$ and $\boldsymbol{b}(t)$:

$$\begin{aligned}\Big\|\sum_{t=\Delta t}^{T\Delta t}\{\boldsymbol{a}'(t)\boldsymbol{b}(t)\}\boldsymbol{b}(t)\Big\|^2 &= 2\sum_{t=\Delta t+\Delta t}^{T\Delta t}\{\boldsymbol{a}'(t)\boldsymbol{b}(t)\}\boldsymbol{b}'(t)\cdot\sum_{j=\Delta t}^{t-\Delta t}\{\boldsymbol{a}'(j)\boldsymbol{b}(j)\}\boldsymbol{b}(j)\\ &\quad + \sum_{t=\Delta t}^{T\Delta t}\{\boldsymbol{a}'(t)\boldsymbol{b}(t)\}^2\{\boldsymbol{b}'(t)\boldsymbol{b}(t)\}.\end{aligned} \tag{15}$$

Consider the first term which appears on the right side of (14). From (8),

$$\sum_{t=t_0}^{t_0+(T-1)\Delta t}\{\boldsymbol{x}'(t)\boldsymbol{r}(t)\}^2 = \frac{1}{K(2-K)}\left(\|\boldsymbol{r}(t_0)\|^2-\|\boldsymbol{r}(t_0+T\Delta t)\|^2\right). \tag{16}$$

Now consider the second term on the right side of (14):

$$\begin{aligned}&\sum_{t=t_0}^{t_0+(T-1)\Delta t}\left[\boldsymbol{x}'(t)\{\boldsymbol{r}(t_0)-\boldsymbol{r}(t)\}\right]^2\\ &\leqslant \sum_{t=t_0}^{t_0+(T-1)\Delta t}\|\boldsymbol{r}(t_0)-\boldsymbol{r}(t)\|^2\\ &= \sum_{t=t_0+\Delta t}^{t_0+(T-1)\Delta t}\Big\|\sum_{s=t_0}^{t-\Delta t}K\{\boldsymbol{r}'(s)\boldsymbol{x}(s)\}\boldsymbol{x}(s)\Big\|^2, \quad \text{from (6)},\\ &\leqslant K^2\sum_{t=t_0+\Delta t}^{t_0+(T-1)\Delta t}\left[\sum_{s=t_0}^{t-\Delta t}\|\boldsymbol{x}(s)\|^2\sum_{s=t_0}^{t-\Delta t}\{\boldsymbol{r}'(s)\boldsymbol{x}(s)\}^2\right],\end{aligned}$$

by Schwarz's inequality,

$$= K^2 \sum_{t=t_0+\Delta t}^{t_0+(T-1)\Delta t} \left[\frac{t-t_0}{\Delta t} \sum_{s=t_0}^{t-\Delta t} \{r'(s)x(s)\}^2 \right],$$

since $\|x(s)\|^2 \equiv 1$,

$$= K^2 \sum_{t=t_0+\Delta t}^{t_0+(T-1)\Delta t} \frac{t-t_0}{\Delta t} \cdot \frac{1}{K(2-K)} \left[\|r(t_0)\|^2 - \|r(t)\|^2 \right],$$

from (8),

$$\leq \frac{K}{2-K} \left[\|r(t_0)\|^2 - \|r(t_0+T\Delta t)\|^2 \right] \sum_{t=t_0+\Delta t}^{t_0+(T-1)\Delta t} \frac{t-t_0}{\Delta t},$$

from (8),

$$= \frac{KT(T-1)}{2(2-K)} \left[\|r(t_0)\|^2 - \|r(t_0+T\Delta t)\|^2 \right], \tag{17}$$

where the final step follows from the identity $\sum_{i=1}^{N} i = N(N+1)/2$.

Finally, consider the third term in the right side of (14):

$$2 \sum_{t=t_0}^{t_0+(T-1)\Delta t} \{x'(t)r(t)\} \left[x'(t)\{r(t_0)-r(t)\} \right]$$

$$= 2K \sum_{t=t_0+\Delta t}^{t_0+(T-1)\Delta t} \{x'(t)r(t)\} x'(t) \sum_{s=t_0}^{t-\Delta t} \{r'(s)x(s)\} x(s),$$

from (6),

$$= K \Big\| \sum_{t=t_0}^{t_0+(T-1)\Delta t} \{x'(t)r(t)\} x(t) \Big\|^2 - K \sum_{t=t_0}^{t_0+(T-1)\Delta t} \{x'(t)r(t)\}^2, \quad \text{from (15)},$$

$$\leq K \sum_{t=t_0}^{t_0+(T-1)\Delta t} \|x(t)\|^2 \sum_{t=t_0+\Delta t}^{t_0+(T-1)\Delta t} \{x'(t)r(t)\}^2 - K \sum_{t=t_0}^{t_0+(T-1)\Delta t} \{x'(t)r(t)\}^2, \quad \text{by Schwarz's inequality},$$

$$= K(T-1) \frac{\left[\|r(t_0)\|^2 - \|r(t_0+T\Delta t)\|^2 \right]}{K(2-K)}, \quad \text{from (8)},$$

$$= \frac{T-1}{2-K} \left[\|r(t_0)\|^2 - \|r(t_0+T\Delta t)\|^2 \right]. \tag{18}$$

On substituting the bounds in (16), (17), and (18) into (14), we obtain

$$\|r(t_0+T\Delta t)\|^2 \leq \|r(t_0)\|^2 \left[1 - \frac{2\alpha KT(2-K)}{2+2K(T-1)+K^2T(T-1)} \right]. \tag{19}^2$$

The above equation is equivalent to the promised exponential bound. If we take b such that

$$b = \frac{1}{2T} \ln \left[1 - \frac{2\alpha KT(2-K)}{2+2K(T-1)+K^2T(T-1)} \right] \tag{20}$$

then

$$\|r(t_0+N\Delta t)\| \leq \begin{cases} \|r(t_0)\|, & \text{for } N \leq T \\ \|r(t_0)\| e^{-b(N-T)}, & \text{for } N > T. \end{cases} \tag{21}$$

Observe that $b > 0$ if $K < 2$.

We have also shown (the proof is omitted) by a rather different method that

$$\|r(t_0+T\Delta t)\| \leq \gamma_0 \|r(t_0)\|, \qquad \text{for all } t_0,$$

where γ_0 is the unique positive root of

$$\left[1 + \alpha KT + \alpha(\alpha T+1)K^2T/2 \right] \gamma = 1 + K^3 \left[\frac{1-\gamma^2}{2K(2-K)} \right]^{1/2} \cdot \left[\frac{T^3(T+1)^2}{4} + \frac{T^2(T+1)(2T+1)}{6} \right]^{1/2}$$

with $0 < \gamma_0 < 1$. The above bound is superior to (19) only for large values of α.

Our results are summarized in the next proposition.

Proposition 1: If $x(t)$ satisfies the mixing condition (10) and $r(t)$ satisfies (6), then

$$\|r(t_0+T\Delta t)\| \leq B \|r(t_0)\|, \qquad \text{for all } t_0,$$

where B is the smaller of the quantities

$$\left[1 - \frac{2\alpha KT(2-K)}{2+2K(T-1)+K^2T(T-1)} \right]^{1/2}$$

and γ_0. Thus, for any $N \geq 0$,

$$\|r(t_0+N\Delta t)\| \leq a e^{-bN} \|r(t_0)\|$$

where $b = -(\ln B)/T$ and $a = e^{bT}$.

If we pass to the continuous limit in (19), that is, let $K \to 0$ and $T \to \infty$ in such a way that $KT = \text{constant} = K'T'$ and $T\Delta t = \text{constant} = T'$ then we find that

$$\|r(t_0+T')\|^2 \leq \|r(t_0)\|^2 \left[1 - \frac{4\alpha K'T'}{2+2K'T'+K'^2T'^2} \right] \tag{22}$$

which is identical to Sondhi and Mitra's [1, (26) and (27)].

B. Dependence of the Upper Bound on the Loop Gain K

We now examine the manner in which b, which indicates the rate of convergence, depends on K. We do not consider the dependence on α and T, since these parameters are inherent to the input process and not subject to control. We consider the case where

$$b = -\frac{1}{2T} \ln \left[1 - \frac{2\alpha KT(2-K)}{2+2K(T-1)+K^2T(T-1)} \right]. \tag{20}$$

[2] We observe that if $\|x(t)\|$ is, as in (4'), uniformly bounded by L insteady of being normalized to unity, as has been assumed throughout, then the above procedure yields

$$\|r(t_0+T\Delta t)\|^2 \leq \|r(t_0)\|^2 \left(1 - \frac{2\alpha KT(2-KL^2)}{2+2K(L^2T-1)+K^2L^4T(T-1)} \right).$$

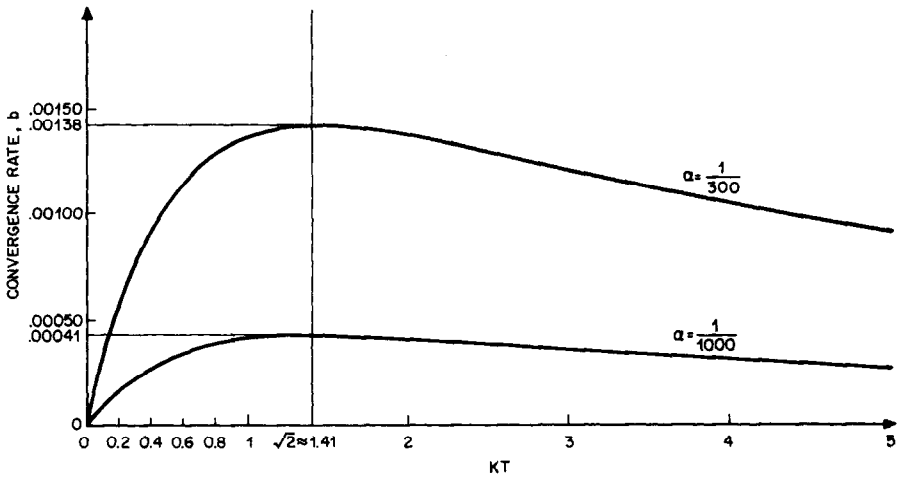

Fig. 2. Convergence rate derived from upper bound (Section II-A). $b=-(1/2T)\ln[1-4\alpha KT/(2+2KT+K^2T^2)]$.

If K is small then $2-K\approx 2$ and using the approximation $\ln(1-a)\approx -a$ for small values of a, we obtain

$$b\approx\frac{1}{2T}\frac{4\alpha KT}{2+2K(T-1)+K^2T(T-1)}\approx\alpha K. \tag{23}$$

If K is large, that is $K\to 1$, then we ignore terms of order less than two in T (recall that $T\geqslant n$, the dimension of $x(t)$, typically a large number) and we obtain from (20)

$$b\approx-\frac{1}{2T}\ln\left[1-\frac{2\alpha KT(2-K)}{K^2T^2}\right]\approx\frac{\alpha(2-K)}{KT^2}\approx\frac{\alpha}{KT^2}. \tag{24}$$

Observe that $\alpha/T^2\leqslant 1/n^3$, a very small number.

We see that in the exponential bound the rate of convergence b increases linearly with K for small K, and is inversely proportional to K as K approaches 1. As K approaches 1, the exponent in the exponential bound rapidly approaches the very small number α/T^2. A graph of $b(K)$ for certain values of α and T which demonstrates this behavior is given in Fig. 2.

The optimum value of K, i.e., that value of K for which $\|r(t)\|$ decreases most rapidly, as suggested by our upper bound, is easily calculated to be (after setting $db/dK=0$ and solving)

$$K=\frac{\sqrt{2T^2-1}-1}{T^2-1}\approx\frac{\sqrt{2}}{T}. \tag{25}$$

Since T is typically large ($T\geqslant n$) we see that the optimum value of K is rather small compared to one.

C. Discussion of the Optimum Value of the Loop Gain K

The optimum value just calculated from our upper bound is quite different from the best value of K obtained from 'myopic' optimization: examining (8) we see that $\|r(t)\|-\|r(t+\Delta t)\|$ is maximal when $K=1$. This is not surprising since maximizing $|\Delta\|r(t)\||$ after each interval of length Δt may not maximize the change in norm over a collection of intervals.

Another notion which leads to an erroneous 'optimum' value of K is the method of averaging. This involves taking (6), $\Delta r(t)=-Kx(t)x'(t)r(t)$, and assuming that $r(t)$ behaves something like the solution to

$$\Delta r(t)=-KAr(t) \tag{26}$$

where A is the $n\times n$ matrix which is the expected value of $x(t)x'(t)$. We can show that the method of averaging is not very useful for large K, or even for values of K for which our upper bound on the convergence rate is optimal. We note parenthetically that we will use something similar to the averaging argument for the case of small K in the following section on lower bounds.

We also observe that for a particular process the true optimal value of K (that value which makes for the fastest decrease in $\|r(t)\|$) may be quite different from the value given in (25). In fact, we have some (rather pathological) examples for which the optimal value of K is indeed 1.[3] However, for many processes $x(t)$, we are in a position to indicate an interval in which the true optimum lies. We do this by finding exponentially decreasing lower bounds on $\|r(t)\|$ where the exponents have the same behavior with K as our upper bound; that is, we find lower bounds whose exponent increases linearly with K for small K and (for a wide class of processes) decreases as $1/K$ as K approaches 1. As shown in Fig. 3, this will give bounds on the range of the true optimal value of K for a given process $x(t)$.

III. Lower Bounds

We obtain our first lower bound on $\|r(t)\|$ directly from (6):

$$\|r(t+\Delta t)\|=\|r(t)-K(r'(t)x(t))x(t)\|$$
$$\geqslant\|r(t)\|-K\|(r'(t)x(t))x(t)\| \tag{27}$$
$$\geqslant\|r(t)\|(1-K). \tag{28}$$

[3]One excellent example is a process $x(t)$ where $x(t)$ is perpendicular to $x(t-j\Delta t)$ for $j=1,2,\cdots,n-1$. In this example we have $\alpha=1/n$, $T=n$. If we take $K=1$ here, then $r(t_0+n\Delta t)=0$! Any K smaller than unity will not perform as well.

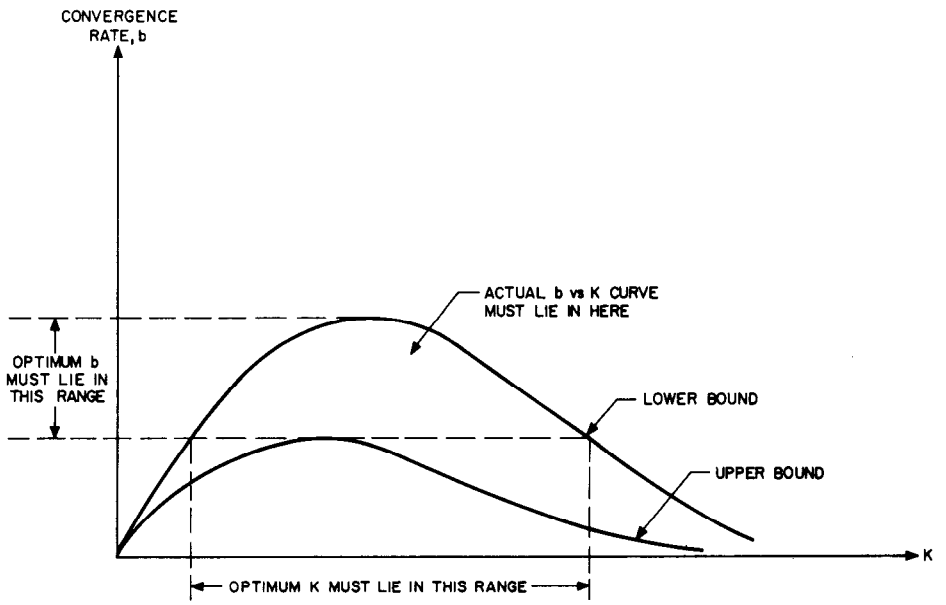

Fig. 3. Sketch illustrating role of upper and lower bounds in determining range of values of K of particular interest for given input process.

For small K we see that this is equivalent to

$$\|r(t_0+N\Delta t)\| > e^{N\log(1-K)}\|r(t_0)\|, \tag{29}$$

i.e., we have an exponential lower bound whose exponent increases linearly with K for small K.

Unfortunately, this simple lower bound leaves much to be desired since, as detailed in the discussion at the end of Section III-A, it does not have anything like the behavior of the upper bound in Proposition 1, Section II-A, for large values of K. For extremely small K we obtain below a sharper lower bound via the mixing condition. For larger values of K we obtain a lower bound summarized in Proposition 2, Section III-C.

A. Lower Bound for Very Small K

Intuitively, if K is very small, then $r(t)$ does not change very rapidly, so we suspect $r(t)\approx r(t_0)$ for $t-t_0 \leqslant T\Delta t$, where T is as in the mixing condition (10). We then have

$$\|r(t_0+T\Delta t)\|^2 = \|r(t_0)\|^2 - K(2-K) \cdot \sum_{t=t_0}^{t_0+(T-1)\Delta t} \{r'(t)x(t)\}^2, \quad \text{from (8)},$$

$$\approx \|r(t_0)\|^2 - 2K \sum_{t=t_0}^{t_0+(T-1)\Delta t} \{r'(t_0)x(t)\}^2, \tag{30}$$

$$\geqslant \|r(t_0)\|^2 - 2KT\|r(t_0)\|^2\{1-(n-1)\alpha\}, \tag{31}$$

$$= \|r(t_0)\|^2[1-2KT\{1-(n-1)\alpha\}]. \tag{32}$$

Equation (30) follows from replacing $2-K$ by 2 and $r(t)$ by $r(t_0)$. By using (28) and the bound in Proposition 1, it is easy to show that the errors incurred in (30) are of order K^2. Inequality (31) comes from the fact that the largest eigenvalue of

$$\left\{ \sum_{t=t_0}^{t_0+(T-1)\Delta t} x(t)x'(t) \right\} / T$$

is at most $1-(n-1)\alpha$. From (32) we have

$$\|r(t_0+N\Delta t)\| \geqslant \|r(t_0)\| e^{-c(N-T)} \tag{33}$$

where

$$c \triangleq -\frac{1}{2T}\ln[1-2KT\{1-(n-1)\alpha\}] \approx K\{1-(n-1)\alpha\}. \tag{34}$$

The above is better than the bound in (29) which has $c\approx K$. In fact, if $\alpha\approx 1/n$ (the maximum possible value), (34) gives $c\approx K/n$, which is the same as the upper bound (23). This shows that the bound in (34) is the best possible when all that is known about the input is that it is mixing.

We now have the best possible lower bound for small K, (33), and we also have a lower bound for all K, (29). However, the exponent in the exponential bound implied by (29) grows monotonically with K, in sharp contrast with our upper bound where it decreases as $1/K$ for large K. In order to establish this behavior for the lower bound we will have to examine the convergence process in greater detail.

B. Geometrical Preliminaries: K Not Small

It will be beneficial to give the reader a flavor of the final result of this section which is developed almost exclusively from geometrical arguments. Consider the plane formed by the generic $x(t)$ and $r(t)$ vectors, see Fig. 4. Provided K is large relative to the rate at which the vectors $x(t)$ may change, we show that there exist two regions in the plane, B_1 and B_2, with the following important properties. Region B_2 acts as a 'trap' region in the

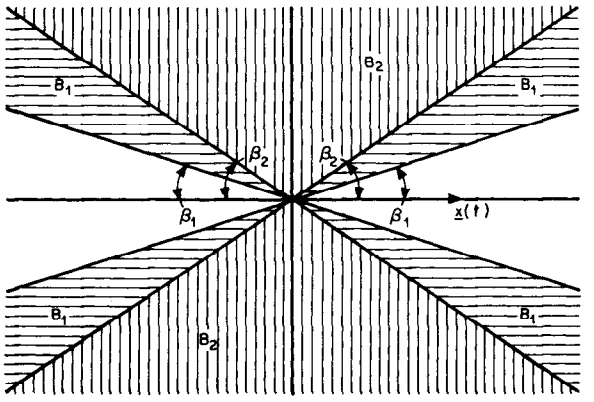

Fig. 4. Regions B_1 and B_2 in the $(x(t),r(t))$ plane. See Section III-B on lower bound.

sense that if $\{x(t),r(t)\}$ occurs in B_2 then so does $\{x(t+\Delta t),r(t+\Delta t)\}$ and consequently all subsequent such pairs also. Region B_1 acts as a 'drift' region in the sense that if $\{x(t),r(t)\}$ occurs in B_1, then $r(t+\Delta t)$ lies closer to the region B_2 in the $\{x(t+\Delta t),r(t+\Delta t)\}$ plane. Eventually the pairs of vectors are guaranteed to 'drift' into the trap region B_2. We do not make any claims regarding the subsequent behavior of the system if $\{x(t),r(t)\}$ does not lie in either B_1 or B_2. Note that the regions B_1 and B_2 are completely defined by the angles β_1 and β_2, $0 \leqslant \beta_1 < \beta_2 < \pi/2$:

$$\{x(t),r(t)\} \in B_1 \Leftrightarrow \cos\beta_2 \leqslant \frac{|x'(t)r(t)|}{\|r(t)\|} \leqslant \cos\beta_1, \quad (35)$$

$$\{x(t),r(t)\} \in B_2 \Leftrightarrow 0 \leqslant \frac{|x'(t)r(t)|}{\|r(t)\|} \leqslant \cos\beta_2. \quad (36)$$

Let θ_t be the angle between $x(t)$ and $r(t)$, and let ϕ_t be the angle between $x(t)$ and $r(t+\Delta t)$ in the plane formed by the $x(t),r(t)$ vectors. (Note that $r(t+\Delta t)$ is in the $\{x(t),r(t)\}$ plane as $r(t+\Delta t)=r(t)-K\{r'(t)x(t)\}x(t)$ is a linear combination of the vectors in the plane.) For most of this section our only contact with the dynamics of the $\{r(t)\}$ process will be through the following geometrical statement relating ϕ_t and θ_t:

$$\tan\phi_t = \frac{\tan\theta_t}{1-K}. \quad (37)$$

This is clear from Fig. 5, and it may also be easily proved analytically from (6). The lower bound that we derive (for K not small) requires the assumption that

$$\|x(t+\Delta t)-x(t)\| < \sqrt{2}, \qquad \text{for all } t. \quad (38)$$

More precisely we require an angle δ, $0<\delta<\pi/2$, such that

Assumption:

$$\|x(t+\Delta t)-x(t)\| \leqslant \sqrt{2(1-\cos\delta)} < \sqrt{2}, \qquad \text{for all } t \quad (39)$$

or, equivalently,

$$x'(t+\Delta t)x(t) \geqslant \cos\delta > 0, \qquad \text{for all } t. \quad (40)$$

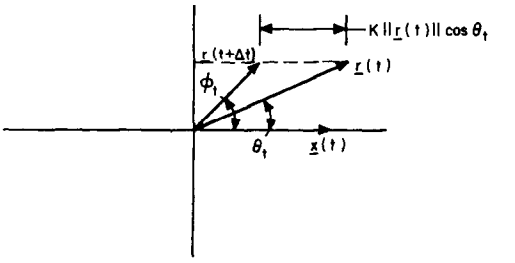

Fig. 5. $(x(t),r(t))$ plane. See Section III-B on lower bound.

Thus the angle between $x(t+\Delta t)$ and $x(t)$ is bounded by a number $\delta<\pi/2$, for all $t \geqslant t_0$. Another way of interpreting this condition is that the 'velocity' of the input process is bounded by a finite number (see [1] for a similar restriction). This assumption is somewhat related to the mixing condition. If δ is small then $x(t)$ cannot move very rapidly, and so cannot "mix" well in a short period of time. This means that if δ is small then we cannot have both T small and α large in the mixing condition, see (10) in Section I-D.

Relying only on the assumption on $x(t)$, (39), we claim that

$$|\phi_t - \theta_{t+\Delta t}| \leqslant \delta, \quad (41)$$

that is, the angle between $x(t+\Delta t)$ and $r(t+\Delta t)$ is within δ of the angle between $x(t)$ and $r(t+\Delta t)$. This is so because the angle between $x(t+\Delta t)$ and $x(t)$ is, by (39), bounded by δ.

From (41) we conclude that

$$|\pi/2-\theta_{t+\Delta t}| \leqslant |\pi/2-\phi_t| + \delta. \quad (42)$$

Let us now try to find angles θ_t, for given K and δ, such that (compare with (41)),

$$|\phi_t - \theta_t| = \delta, \quad (43)$$

that is, the angle between $r(t)$ and $r(t+\Delta t)$ is δ. It is understood that ϕ_t is related to θ_t through (37). The solutions θ_t of (43) will prove important for our bounds.

We now show that if K is large or δ small then there are exactly two solutions, β_1 and β_2, for angles θ_t in the interval $[0,\pi/2]$ which satisfy the following pairs of equations:

$$\tan\phi_t = \frac{\tan\theta_t}{1-K}, \qquad 0 \leqslant \phi_t \leqslant \pi/2. \quad (44)$$

$$\phi_t - \theta_t = \delta. \quad (45)$$

If we call the solutions to these equations β, and substitute (45) into (44), we obtain the single equation

$$\tan(\beta+\delta) = \frac{\tan\beta}{1-K}. \quad (46)$$

Note parenthetically that $\delta<\beta+\delta<\pi/2$ since $\tan(\beta+\delta)$ and $\tan\beta$ have the same sign. On expanding the left side of (46),

$$\tan(\beta+\delta) = \frac{\tan\beta+\tan\delta}{1-\tan\beta\tan\delta},$$

we observe that (46) is a quadratic in $\tan\beta$. The solutions

are

$$\tan\beta_{1,2}=\frac{K\mp\sqrt{K^2-4(1-K)\tan^2\delta}}{2\tan\delta}. \tag{47}$$

For solutions to exist it is necessary and sufficient that the following assumption relating K and δ be valid.

Assumption:

$$\frac{K}{\sqrt{1-K}}\geqslant 2\tan\delta. \tag{48}$$

We see that as $\tan\delta\to\infty$ ($\delta\to\pi/2$) we need $K\to 1$ for the solutions to exist.

At this stage then we have that if K is large or δ small, i.e., (48) is valid, then two solutions β_1 and β_2 of (47) exist. We have already seen that $0<\beta_1<\beta_2<\pi/2$. These angles, β_1 and β_2, are used to define the regions B_1 and B_2 as in Fig. 4 and (35) and (36). Now

$$\tan(\phi_t-\theta_t)=\frac{K\tan\theta_t}{1-K+\tan^2\theta_t}. \tag{49}$$

An elementary calculations shows that this expression is strictly increasing for $0\leqslant\tan\theta_t<\sqrt{1-K}$, and is strictly decreasing for $\sqrt{1-K}<\tan\theta_t<\infty$. Recalling that if $\theta_t=\beta_1$ or β_2 then $\tan(\phi_t-\theta_t)=\tan\delta$, we have

$$\text{if } \beta_1<\theta_t<\beta_2, \text{ then } \tan\delta<\tan(\phi_t-\theta_t),$$
$$\text{i.e., } \delta<\phi_t-\theta_t. \tag{50}$$

From (42), for $\beta_1<\theta_t<\beta_2$ (recall that this implies $\phi_t<\pi/2$),

$$\begin{aligned}|\pi/2-\theta_{t+\Delta t}|&\leqslant\pi/2-\phi_t+\delta\\&<\pi/2-\theta_t=|\pi/2-\theta_t|.\end{aligned} \tag{51}$$

Considering the three cases $-\beta_2<\theta_t<-\beta_1$, $\pi-\beta_1<\theta_t<\pi-\beta_2$, and $\pi+\beta_1<\theta_t<\pi+\beta_2$ separately, we find that in each case (51) holds. This can be put concisely in the form of a picture:

> The region B_1, see Fig. 4, acts as a 'drift' region in the following sense. If $\{x(t),r(t)\}$ lies in B_1 in the $\{x(t),r(t)\}$ plane, then $r(t+\Delta t)$ lies closer to the perpendicular to $x(t+\Delta t)$ in the $\{x(t+\Delta t),r(t+\Delta t)\}$ plane.

(52)

We now examine the region B_2. We claim that

$$\begin{aligned}&\text{if } |\pi/2-\theta_t|<\pi/2-\beta_2,\\&\text{then } |\pi/2-\theta_{t+\Delta t}|\leqslant\pi/2-\beta_2.\end{aligned} \tag{53}$$

For

$$\begin{aligned}|\pi/2-\theta_t|\leqslant\pi/2-\beta_2&\Rightarrow|\tan\theta_t|\geqslant\tan\beta_2\\&\Rightarrow\frac{|\tan\theta_t|}{1-K}\geqslant\frac{\tan\beta_2}{1-K}\\&\Rightarrow|\tan\phi_t|\geqslant\tan(\beta_2+\delta),\quad\text{from (37) and (46),}\\&\Rightarrow|\pi/2-\phi|\leqslant\pi/2-(\beta_2+\delta)\\&\Rightarrow|\pi/2-\theta_{t+\Delta t}|\leqslant\pi/2-\beta_2,\quad\text{from (43).}\end{aligned}$$

The results in (53) can also be stated concisely in geometrical terms:

> If $r(t)$ is an element of B_2 in the $\{x(t),r(t)\}$ plane, then $r(t+\Delta t)$ also lies in the corresponding region in the $\{x(t+\Delta t),r(t+\Delta t)\}$ plane.

(54)

The statements (52) and (54) concerning the regions B_1 and B_2 summarize the results of this section. These results are contingent upon the assumptions that K is large or δ small, as in (39) and (48), for only then do these regions exist.

C. Analytic Bounds: K Not Small

We are now ready to give our lower bound. If $r(t_0)$ lies in the region B_2 of Fig. 4, then all succeeding $r(t)$ will also lie in that region. Thus if $|\cos\theta_{t_0}|\leqslant\cos\beta_2$ then $|\cos\theta_t|\leqslant\cos\beta_2$ for all $t\geqslant t_0$. This leads directly to

$$\begin{aligned}\|r(t+\Delta t)\|^2&=\|r(t)\|^2\{1-K(2-K)\cos^2\theta_t\},\quad\text{from (8)}\\&\geqslant\|r(t)\|^2\{1-K(2-K)\cos^2\beta_2\}\end{aligned} \tag{55}$$

which in turn implies

$$\|r(t_0+N\Delta t)\|\geqslant\|r(t_0)\|e^{-cN} \tag{56}$$

where

$$c=-\frac{1}{2}\ln\{1-K(2-K)\cos^2\beta_2\}. \tag{57}$$

If $r(t_0)$ is in region B_1 of Fig. 4 then we know from (51) that $\cos^2\theta_{t_0+\Delta t}<\cos^2\theta_{t_0}$. This leads to a bound indentical to (56) and (57) with β_2 replaced by θ_{t_0}.

Actually, we can do better than this. If $r(t_0)$ is in B_1 then there is no loss of generality in assuming $\beta_1<\theta_{t_0}<\beta_2$. We see then that

$$\begin{aligned}\tan\theta_{t+\Delta t}&\geqslant\tan(\phi_t-\delta),\quad\text{from (42),}\\&=(\tan\phi_t-\tan\delta)/\{1+\tan\phi_t\tan\delta\},\\&=\frac{\tan\theta_t-(1-K)\tan\delta}{\tan\theta_t\tan\delta+(1-K)},\quad\text{from (37).}\end{aligned} \tag{58}$$

This provides the basis for a recursive lower bound on θ_t for $t\geqslant t_0$. From (8) we directly obtain the following lower bound:

$$\|r(t_0+N\Delta t)\|\geqslant\|r(t_0)\|\prod_{j=t_0}^{t_0+N-1}\{1-K(2-K)\cos^2 a_{j-t_0}\}^{1/2} \tag{59}$$

where the angles $\{a_j\}$ satisfy the recursion

$$\tan a_{j+1}=\frac{\tan a_j-(1-K)\tan\delta}{\tan a_j\tan\delta+(1-K)},\qquad a_0=\theta_{t_0}. \tag{60}$$

We know from previous considerations that $a_{j+1}>a_j$ and $a_j\to\beta_2$. It is difficult to solve (60) in closed form, but numerical answers can be easily obtained for any θ_{t_0}, K and δ. We summarize the lower bound in the following.

Proposition 2: Suppose that the input vectors $x(t)$ are such that the angle between $x(t+\Delta t)$ and $x(t)$ is bounded

by a number $\delta<\pi/2$, i.e.,

$$\|x(t+\Delta t)-x(t)\| \leqslant \sqrt{2(1-\cos\delta)} < \sqrt{2}, \qquad \text{for all } t \geqslant t_0,$$

and that K is sufficiently large so that

$$\frac{K}{\sqrt{1-K}} \geqslant 2\tan\delta.$$

Equation (47) gives the two solutions β_1 and β_2, $0<\beta_1<\beta_2<\pi/2$, to the equation

$$\tan(\beta+\delta)=\tan\beta/(1-K).$$

The angles β_1 and β_2 define the regions B_1 and B_2 as in (35) and (36).

i) If $r(t_0)$ and $x(t_0)$ are aligned such that they lie in B_2 then

$$\|r(t_0+N\Delta t)\| \geqslant \|r(t_0)\|e^{-Nc}, \qquad \text{for all } N \geqslant 0, \quad (61)$$

where $c=-\frac{1}{2}\ln\{1-K(2-K)\cos^2\beta_2\}$.

ii) If $r(t_0)$ and $x(t_0)$ are aligned such that they lie in B_1 then

$$\|r(t_0+N\Delta t)\| \geqslant \|r(t_0)\| \prod_{j=t_0}^{t_0+N-1} \{1-K(2-K)\cos^2 a_{j-t_0}\}^{1/2}, \qquad \text{for all } N \geqslant 1, \quad (62)$$

where

$$\tan a_{j+1}=\frac{\tan a_j-(1-K)\tan\delta}{\tan a_j\tan\delta+(1-K)},$$

$$a_0=\cos^{-1}\frac{r(t_0)'x(t_0)}{\|r(t_0)\|}.$$

In particular, $\lim_{j\to\infty} a_j=\beta_2$, i.e., eventually $\{x(t),r(t)\}$ enter the "trap" region B_2.

D. The Continuous Limit

The special case of (59) where $\Delta t\to 0$ (i.e., $K\to 0$ and $\delta\to 0$ in such a way that $K/\delta=$constant) can be solved analytically, yielding the same result as Sondhi and Mitra's [1, eq. (45)]. The details are somewhat cumbersome and are omitted.

E. Dependence of the Lower Bound on the Loop Gain K

We now analyze the behavior of the lower bound as a function of K for large K. We assume $r(t_0)$ is in B_2 since we know that the system tends to this situation for $\{x(t_0),r(t_0)\}$ in B_1 or B_2. From the expression for $\tan\beta_2$ in (47) we see that if $(\tan\delta)/K$ is small, then $\tan\beta_2\approx K/\tan\delta$, or $\pi/2-\beta_2\approx(\tan\delta)/K$; hence $\cos\beta_2\approx(\tan\delta)/K$. Thus in (57)

$$\begin{aligned} c &= -\frac{1}{2}\ln\{1-K(2-K)\cos^2\beta_2\} \\ &\approx -\frac{1}{2}\ln\left\{1-K(2-K)\frac{\tan^2\delta}{K^2}\right\} \\ &\approx \frac{(2-K)\tan^2\delta}{2K}. \end{aligned}$$

As K increases we see that c decreases something like $1/K$. Thus the qualitative similarity of the upper and lower bounds is established. If we let $\Delta t\to 0$ (i.e., $K\to 0$, $\delta\to 0$, $\delta/K=$constant) we obtain $c\approx\tan^2\delta/K$, Sondhi and Mitra's result [1, eq. (40)].

IV. Noise

A. The Forced Equations

We consider the effects of the forcing term $u(t)$ in the equation

$$\Delta r(t)=-K\{r'(t)x(t)\}x(t)+u(t). \qquad (7)$$

The term $u(t)$, an n-vector, can be used to represent the effects of departures from the idealized problem described by the homogeneous version of (7), such as when noise is present in the return signal or the unknown coefficients are varying with time. We will show now that if $\|u(t)\|$ is bounded, or equivalently has a bounded mean over intervals of length T, then the residual error $\|r(t)\|$ remains bounded as $t\to\infty$. Subsequently, by appropriately identifying the forcing term $u(t)$, we will obtain estimates of the effects of the departures from the idealized problem.

Equation (7) may be rewritten as

$$r(t+\Delta t)=[I-Kx(t)x'(t)]r(t)+u(t), \qquad t=t_0,t_0+\Delta t,\cdots. \quad (63)$$

As is well-known [31], there exists a formal solution to (64) in terms of the fundamental matrix $Y(t,t_0)$, $t_0\leqslant t$:

$$r(t)=Y(t,t_0)r(t_0)+\sum_{j=t_0+\Delta t}^{t} Y(t,j)u(j-\Delta t), \qquad (64)$$

Our upper bound developed in Section II-A and summarized in Proposition 1 translates to the following bound in the fundamental matrix:

$$\|Y(t,t_0)\|=e^{-jbT} \text{ when } jT\Delta t \leqslant t-t_0<(j+1)T\Delta t. \quad (65)$$

Assume at this stage that all that is known about the forcing terms $u(t)$ is that the time average of its norm over any T samples is bounded, i.e.,

$$\frac{1}{T}\sum_{j=t}^{t+(T-1)\Delta t}\|u(j)\| \leqslant U, \qquad \text{for all } t \geqslant t_0. \quad (66)$$

Then as $t\to\infty$ we obtain, from (64),

$$\|r(\infty)\| \leqslant UT\sum_{j=0}^{\infty} e^{-jbT}=\frac{UT}{1-e^{-bT}}. \qquad (67)$$

Alternatively if $\|u(t)\|$ is bounded by $\bar{u}$ for all $t\geqslant t_0$, then $U\leqslant\bar{u}$ so that

$$\|r(\infty)\| \leqslant \frac{\bar{u}T}{1-e^{-bT}}. \qquad (68)$$

B. Noise in the Measured Signal

We apply the result in (67) and (68) to some special cases. Suppose that there is noise, or errors in observation,

in the observed signal $z(t)$ which appears in (1). Specifically, suppose that instead of (1) we have

$$z(t)=\boldsymbol{h}'\boldsymbol{x}(t)+s(t) \tag{69}$$

where $s(t)$ is an undesirable noise signal. We take the approach that not much is known about $s(t)$ except that it is bounded. When we follow the effects of this signal through we see that (7) holds with

$$\boldsymbol{u}(t)=-Ks(t)\boldsymbol{x}(t). \tag{70}$$

If

$$\frac{1}{T}\sum_{j=t}^{t+(T-1)\Delta t} s^2(j)\leqslant S^2, \qquad \text{for all } t, \tag{71}$$

then Schwarz's inequality gives

$$\frac{1}{T}\sum_{j=t}^{t+(T-1)\Delta t} \|\boldsymbol{u}(j)\|\leqslant KS \tag{72}$$

which may be used in (67). We have then that

$$\|\boldsymbol{r}(\infty)\|\leqslant \frac{KST}{1-e^{-bT}}. \tag{73}$$

Recall from Section 2-B that b is proportional to K for small K. Thus the bound for the residual error in (73) is independent of K for small K. Also, for K approaching 1, b is approximately α/KT^2 (see (24)), a small number; hence the bound in (73) is proportional to K^2 as K approaches one.

C. Variations in the Coefficient Vector h

Suppose now that the vector $\boldsymbol{h}$ in (1) is not time invariant. Then

$$\begin{aligned}\Delta\boldsymbol{r}(t)&=\Delta\boldsymbol{h}(t)-\Delta\hat{\boldsymbol{h}}(t)\\&=\Delta\boldsymbol{h}(t)-K\{\boldsymbol{r}'(t)\boldsymbol{x}(t)\}\boldsymbol{x}(t).\end{aligned} \tag{74}$$

We see that we may identify the forcing term $\boldsymbol{u}(t)$ with $\Delta\boldsymbol{h}(t)$, and assuming

$$\|\Delta\boldsymbol{h}(t)\|\leqslant H, \qquad \text{for all } t, \tag{75}$$

we find that (68) yields

$$\|\boldsymbol{r}(\infty)\|\leqslant \frac{HT}{1-e^{-bT}}. \tag{76}$$

Thus if $\boldsymbol{h}$ changes slowly with the time the residual error will be small.

An interesting facet of the bound (76) is that it is minimized with respect to K at a value of K which is identical to the value of K which maximizes the rate of convergence of the upper bound derived for the idealized problem. We saw in (25) that this optimum value of K is given by $K\approx\sqrt{2}\,/T$.

V. Effects of Noise and Errors Arising from the Digital Implementation: Mixing and Nonmixing Input

Here we consider the performance of the filter under various departures from the idealized model. We use the language of errors introduced by the implementation. However, with the proper identification of the error terms $\epsilon^{(1)}(t)$ and $\epsilon^{(2)}(t)$ below, the effects of noise in the signals for instance are estimated. The noise can be more general than that considered in Section IV-B (see (83) below) since it may exist in the input vectors $\boldsymbol{x}(t)$ as well as in $z(t)$, although in the simple case the results below are not as sharp.

A. Nonmixing Inputs

If $\boldsymbol{x}(t)$ is not mixing, then we do not necessarily expect the misalignment norm $\|\boldsymbol{r}(t)\|$ to decrease to zero; in fact, Appendix I shows that there is no uniform upper bound on the misalignment. We examine below some of the effects on the convergence process for inputs which belongs to a particular class of nonmixing inputs.

We digress here to explain why we might expect the inputs $\boldsymbol{x}(t)$ not to be mixing in many applications. (See Section I-D for the mixing condition.) In many communications-related applications, such as echo cancellation, $\boldsymbol{x}(t)$ is derived from a speech signal. Typically, a bandpass filtered version of the speech signal is passed through a delay line to yield $\boldsymbol{x}(t)$; thus if $S(t)$ is the bandpass filtered speech signal at time t, then

$$\boldsymbol{x}(t)=[S(t),S(t-\Delta t),\cdots,S\{t-(n-1)\Delta t\}]'. \tag{77}$$

Now consider the constant vector $\boldsymbol{d}$ where

$$\boldsymbol{d}=[\cos(\omega n\Delta t),\cos\{\omega(n-1)\Delta t\},\cdots,\cos(\omega\Delta t)]'. \tag{78}$$

We see that $\boldsymbol{d}'\boldsymbol{x}(t)$ is (approximately) proportional to the Fourier coefficients of $S(t)$ at frequency ω. If we take ω well outside the frequency band to which $S(t)$ is limited, then we might expect $\sum_{t=t_0}^{t_0+(T-1)\Delta t}\{\boldsymbol{d}'\boldsymbol{x}(t)\}^2$ to be quite small, especially if n is large. Thus if $S(t)$ is band-limited, we might expect that $\boldsymbol{x}(t)$ mixes very slowly, if at all.

We see below that even in the case that the input process is not mixing, the results that have been obtained so far with the mixing assumption are applicable with only minor modifications. Suppose that there is a subspace P of R^n for which $P\perp\boldsymbol{x}(t)$ for all $t\geqslant t_0$; that is, $\boldsymbol{x}(t)$ has no component in the space P for any $t\geqslant t_0$. This situation does not exhaust all the possibilities that are associated with the condition "$\boldsymbol{x}(t)$ is not mixing;" however, there will be arbitrarily long time intervals for which this situation is approximated arbitrarily well for any nonmixing input. We also assume that $\boldsymbol{x}(t)$ is mixing in the orthogonal complement of P in R^n which we denote by S, i.e., $S=P^{\perp}$.

In the idealized problem the component of $\boldsymbol{r}(t)$ in the space S will converge exponentially to zero, while the component in the space P will remain constant. This is easily seen from (6) by writing

$$\boldsymbol{rs}(t)=\text{component of } \boldsymbol{r}(t) \text{ in } S, \tag{79}$$

$$\boldsymbol{rp}(t)=\text{component of } \boldsymbol{r}(t) \text{ in } P. \tag{80}$$

Equation (6) becomes

$$\Delta\boldsymbol{rs}(t)=-K\{\boldsymbol{rs}'(t)\boldsymbol{x}(t)\}\boldsymbol{x}(t), \tag{81}$$

$$\Delta\boldsymbol{rp}(t)=0, \tag{82}$$

since $\boldsymbol{rp}'(t)\boldsymbol{x}(t)=0$, and hence $\boldsymbol{r}'(t)\boldsymbol{x}(t)=\boldsymbol{rs}'(t)\boldsymbol{x}(t)$.

As (81) concerns only $\boldsymbol{rs}(t)$ and $x(t)$, and $x(t)$ is mixing in the subspace S, the results in Sections II–IV apply. In particular, we know that $\boldsymbol{rs}(t)$ converges exponentially fast to zero. The remaining component $\boldsymbol{rp}(t)$ is completely described by (82). In any case, as $t\to\infty$, $\|\boldsymbol{r}(t)\|$ remains bounded.

The reader may also verify that the important qualitative property of boundedness is preserved even when the noise signal $s(t)$ is present in the measured signal, as in the case considered in Section IV-B, and the input process is nonmixing.

B. Errors due to Digital Implementation

Here we introduce two rather different kinds of errors which arise in digital implementations of the device. Suppose that instead of (2) we have

$$\Delta\hat{\boldsymbol{h}}(t)=K\{z(t)-\hat{z}(t)\}\{x(t)+\epsilon^{(1)}(t)\} \tag{83}$$

where $\hat{z}(t)=\{\hat{\boldsymbol{h}}(t)+\epsilon^{(2)}(t)\}'x(t)$, and $\epsilon^{(1)}(t)$ and $\epsilon^{(2)}(t)$ are random vectors, most likely with small components, which are introduced in the course of implementing the ideal recursion. Fig. 1 illustrates the points at which these errors appear in a schematic of the device. As the effects of the errors differ qualitatively, we find it convenient to make a distinction by referring to $\epsilon^{(1)}(t)$ and $\epsilon^{(2)}(t)$, respectively, as errors of the first and second kind.

Errors of the second kind could arise from a fixed point to floating point conversion in the device [20]; such a conversion would take place if $\hat{\boldsymbol{h}}(t)$ is stored in the fixed point mode, but the multiplications involved in forming $\hat{\boldsymbol{h}}(t)'x(t)$ are effected in floating point. Likewise, errors of the first kind could arise from a floating point to fixed point conversion of $x(t)$ prior to multiplication with $\boldsymbol{K}\{z(t)-\hat{z}(t)\}$.

To see that the model in (83) is more general than that considered in Section IV-B, note that we may identify $s(t)=\epsilon^{(2)}(t)'x(t)$ and make $\epsilon^{(1)}(t)\equiv 0$.

There is yet another, rather important, reason for considering errors of the first kind. In certain implementations, like the COMSAT echo canceller presently being evaluated [19], the signal $x(t)$ (see Fig. 1) is very coarsely quantized prior to multiplication with $K\{z(t)-\hat{z}(t)\}$. The motivation for this is to simplify the design of the multipliers.[4] The errors introduced by the quantization may of course be denoted by $\epsilon^{(1)}(t)$.

Incorporating the errors $\epsilon^{(1)}(t)$ and $\epsilon^{(2)}(t)$ in (6) gives

$$\Delta\boldsymbol{r}(t)=-K\{\boldsymbol{r}'(t)x(t)-\epsilon^{(2)}(t)'x(t)\}\{x(t)+\epsilon^{(1)}(t)\}. \tag{84}$$

It will be assumed throughout that

$$\|\epsilon^{(1)}(t)\|\leq E_1 \qquad \|\epsilon^{(2)}(t)\|\leq E_2, \quad \text{for all } t. \tag{85}$$

C. Qualitative Behavior with Errors of Both Kinds Present

We examine in turn the convergence properties of the solution $\boldsymbol{r}(t)$ of (84) for the cases where $x(t)$ is respectively mixing and nonmixing.

[4]Note that this procedure is not at all the same as using the "nonideal multipliers" described in [32] even though the motivation is the same.

1) Mixing Inputs: Equation (84) may be written thus

$$\Delta\boldsymbol{r}(t)=-K\{\boldsymbol{r}'(t)x(t)\}x(t)-K\{\boldsymbol{r}'(t)x(t)\}\epsilon^{(1)}(t)+\boldsymbol{f}(t) \tag{86}$$

where

$$\boldsymbol{f}(t)=K\{\epsilon^{(2)}(t)'x(t)\}\{x(t)+\epsilon^{(1)}(t)\}. \tag{87}$$

The assumption that the two errors are uniformly bounded implies a uniform *a priori* bound for $\|\boldsymbol{f}(t)\|$. However, this is not the case for the second term in the right side of (86) since a uniform *a priori* bound for $\|\boldsymbol{r}(t)\|$ does not exist. For the same reason (86) is not in a form that has been encountered previously. We need to step back briefly and prove a new result regarding the behavior of solutions of equations like (86).

Lemma: Suppose

$$\Delta\boldsymbol{r}(t)=-Kx(t)x'(t)\boldsymbol{r}(t)+\boldsymbol{m}(t)\boldsymbol{r}(t)+\boldsymbol{f}(t) \tag{88}$$

where $\boldsymbol{m}(t)$ and $\boldsymbol{f}(t)$ are, respectively, $n\times n$ and $n\times 1$ arbitrary sequences such that

$$\|\boldsymbol{m}(t)\|\leq M \text{ and } \|\boldsymbol{f}(t)\|\leq F, \qquad \text{for all } t. \tag{89}$$

Suppose further that $x(t)$ is mixing so that, by Proposition 1 (see also (65)), the fundamental matrix $Y(t,t_0)$ associated with the recursion $\Delta\boldsymbol{r}(t)=-Kx(t)x'(t)\boldsymbol{r}(t)$ satisfies the bound

$$\|Y(t,t_0)\|\leq ae^{-b(t-t_0)/\Delta t}, \qquad \text{for all } t\geq t_0. \tag{90}$$

Then, for all $N\geq 2$,

$$\|\boldsymbol{r}(N\Delta t)\|\leq\|\boldsymbol{r}(0)\|a(1+M)(aM+e^{-b})^{N-1}+\frac{aF}{1-aM-e^{-b}}\{1-(aM+e^{-b})^N\}. \tag{91}$$

In particular, as $N\to\infty$ we have the following result for arbitrary values of $\boldsymbol{r}(0)$:

$$\text{if } aM+e^{-b}<1, \text{ then } \|\boldsymbol{r}(\infty)\|\leq aF/(1-aM-e^{-b}). \tag{92}$$

Proof: A formal solution of (88), see (64), is

$$\boldsymbol{r}(t)=Y(t,0)\boldsymbol{r}(0)+\sum_{j=\Delta t}^{t}Y(t,j)[\boldsymbol{m}(t)\boldsymbol{r}(j-\Delta t)+\boldsymbol{f}(j-\Delta t)], \qquad t=\Delta t,2\Delta t,\cdots \tag{93}$$

Thus

$$\begin{aligned}\|\boldsymbol{r}(N\Delta t)\|&\leq\|Y(N\Delta t,0)\|\,\|\boldsymbol{r}(0)\|+\sum_{j=\Delta t}^{N\Delta t}\|Y(N\Delta t,j)\|[M\|\boldsymbol{r}(j-\Delta t)\|+F],\\&\leq ae^{-bN}\|\boldsymbol{r}(0)\|+aF\sum_{j=\Delta t}^{N\Delta t}e^{-b(N-j/\Delta t)}+aM\sum_{j=\Delta t}^{N\Delta t}e^{-b(N-j/\Delta t)}\|\boldsymbol{r}(j-\Delta t)\|,\end{aligned} \tag{94}$$

i.e.,

$$e^{bN}\|\boldsymbol{r}(N\Delta t)\|\leq\left[a\|\boldsymbol{r}(0)\|+\frac{aFe^{b}(e^{bN}-1)}{e^{b}-1}\right]+aMe^{b}\sum_{j=0}^{(N-1)\Delta t}e^{bj/\Delta}\|\boldsymbol{r}(j)\|. \tag{95}$$

At this stage we need a discrete-time version of Gronwall's lemma [33]:

If μ is a constant and if $y_N \leq \lambda_N + \sum_{j=0}^{N-1} \mu y_j$ for $N = 1,2,\cdots$, then

$$y_N \leq \lambda_N + \mu(\mu+1)^{N-1}\left\{\sum_{j=1}^{N-1} \lambda_j (1+\mu)^{-j} + y_0\right\} \quad \text{for all } N \geq 2. \quad (96)$$

This is easily established by induction, and we will not prove it here.

We make the identification $y_N = e^{bN}\|r(N\Delta t)\|$ and the natural identification for λ_N and μ. After some straightforward manipulations we find that for $N \geq 2$

$$\|r(N\Delta t)\| \leq \|r(0)\| a(e^{-b}+M)(aM+e^{-b})^{N-1} + \frac{aF}{1-aM-e^{-b}}\{1-(aM+e^{-b})^N\}, \quad (97)$$

from which (91) follows. Observe that the right side diverges if $(aM+e^{-b})>1$. If, on the other hand, $(aM+e^{-b})<1$ then $\|r(N\Delta t)\|$ has the asymptotic upper bound given in (92). This concludes the proof of the Lemma. □

Returning to (86), we find that

$$\|\epsilon^{(1)}(t)x'(t)\| \leq E_1 \qquad \|f(t)\| \leq KE_2(1+E_1), \quad \text{for all } t. \quad (98)$$

An application of the Lemma gives

$$\text{if } aKE_1 + e^{-b} < 1, \text{ then } \|r(\infty)\| \leq \frac{aK(1+E_1)E_2}{1-aKE_1-e^{-b}}. \quad (99)$$

This is important. We observe that the condition for boundedness is satisfied if E_1, which bounds the energy in errors of the first kind, is sufficiently small. It might also appear that, regardless of the value of E_1, the condition is satisfied if K is sufficiently small, but this is not so. Closer inspection shows that it is necessary that $E_1 < \alpha T$; if the latter is true and K is sufficiently small then the condition for boundedness is satisfied. In any case, if neither E_1 nor K is small then the condition for boundedness is violated.

Note the qualitative difference between the two types of errors: errors of the second kind affect the bound quantitatively, in fact linearly, but the condition for boundedness is independent of E_2 while depending on E_1. Errors of the first kind have more influence in determining the qualitative behavior of the device.

2) Nonmixing Inputs: We now consider the case where $x(t)$ is restricted to a subspace S wherein it is mixing. Call $w(t)$ the projection of $\epsilon^{(1)}(t)$ on S, and call $v(t)$ the projection of $\epsilon^{(1)}(t)$ on P, the orthogonal complement of S. Then we may write (84) as

$$\Delta r(t) = -\{r'(t)x(t)\}x(t) - K\{r'(t)x(t)\}w(t) + u(t) + K\{\epsilon^{(2)}(t)'x(t) - r'(t)x(t)\}v(t) \quad (100)$$

where

$$u(t) = K\{\epsilon^{(2)}(t)'x(t)\}\{x(t)+w(t)\}.$$

Observe that the vectors $\{u(t)\}$ are restricted to the subspace S.

We may obtain separate equations for $rs(t)$ and $rp(t)$, as in Section V-A:

$$\Delta rs(t) = -K\{rs'(t)x(t)\}x(t) - K\{rs'(t)x(t)\}w(t) + u(t) \quad (101)$$

$$\Delta rp(t) = K\{\epsilon^{(2)}(t)'x(t) - rs'(t)x(t)\}v(t). \quad (102)$$

Note that the recursion for rs does not depend on rp; in contrast, the recursion for rp depends on rs but not on rp.

Consider (101) first. Observe that the equation is in the form of the equation investigated in the Lemma—$rs(t)$ is restricted to the subspace S, and $x(t)$ is mixing in the subspace S. Application of the Lemma shows that

$$\text{if } \|w(t)\| \leq W(\leq E_1) \qquad \text{for all } t, \quad (103)$$

and

$$\text{if } aKW + e^{-b} < 1, \text{ then } \|rs(\infty)\| \leq \frac{aK(1+W)E_2}{1-aKW-e^{-b}}. \quad (104)$$

In brief, all the results in Section V-C1 on $r(t)$ for mixing input apply here to $rs(t)$.

Now consider (102). The point to note is that it is qualitatively different from (101). Equation (101) contains a stabilizing term in the right side which through the mixing mechanism acts to reduce $\|rs(t)\|$ whenever the latter is large. As the right side of (102) is independent of $rp(t)$, no such mechanism exists to stabilize $\|rp(t)\|$.

The vector $rp(t)$ performs a random walk in the subspace P with random step size and direction. Even if $rs(t)$ is bounded, the right side of (102) may have a nonzero mean since $\epsilon^{(2)}(t)$ is random; in this case $\|rp(t)\|$ will grow linearly with t. Now even if the expectation of the right side of (102) is zero, the norm of $rp(t)$ will grow like $\sqrt{t}$ because of random fluctuations. We thus conclude that, even if $\|\epsilon^{(1)}(t)\|$ and $\|\epsilon^{(2)}(t)\|$ have arbitrarily small bounds, the quantity $\|rp(t)\|$, and consequently also $\|r(t)\|$, will become arbitrarily large after a sufficiently long period of time! This is in sharp contrast with our results in Section V-C 1 for mixing inputs.

It is interesting that the two hardware implementations that we are acquainted with [19], [20] have on occasions demonstrated such unbounded behavior.

We should point out that the quantity $r'(t)xt)$, which is of interest in many applications (in the echo cancellation application, the uncancelled echo is given by $z(t)-\hat{z}(t) = r'(t)x(t) - \epsilon^{(2)}(t)'x(t)$), is uniformly bounded simply because $r'(t)x(t) = rs'(t)x(t)$.

We should also note that the well-known technique of introducing 'leakage' in the adaptation can stabilize the filter at the cost of introducing a residual misalignment error. We omit an analysis of the effects of leakage on the adaptation because a related analysis for the continuous-time algorithm may be found in [1].

D. Both Kinds of Errors Not Simultaneously Present

It is of additional interest that, as we show now, neither one of the two kinds of errors is by itself sufficient to

bring about unbounded growth if its energy is bounded by a small number and the loop gain is small.

1) $\epsilon^{(1)}(t)\equiv 0$*:* The above statement is easily substantiated when $\epsilon^{(2)}(t)$ is the only source of error. Equation (84) then reduces to

$$\Delta \boldsymbol{r}(t) = -K\{\boldsymbol{r}'(t)\boldsymbol{x}(t)\}\boldsymbol{x}(t) + K\{\epsilon^{(2)}(t)'\boldsymbol{x}(t)\}\boldsymbol{x}(t). \tag{105}$$

This immediately gives

$$\Delta \boldsymbol{rs}(t) = -K\{\boldsymbol{rs}(t)'\boldsymbol{x}(t)\}\boldsymbol{x}(t) + K\{\epsilon^{(2)}(t)'\boldsymbol{x}(t)\}\boldsymbol{x}(t), \tag{106}$$

$$\Delta \boldsymbol{rp}(t) = 0. \tag{107}$$

Equation (106) is in a form to which the results of Section IV and the Lemma in Section V-C1 apply. We may conclude that $\|\boldsymbol{rs}(t)\|$ is bounded without making any special restrictions on the loop gain or on the energy of $\epsilon^{(2)}(t)$. Clearly $\|\boldsymbol{rp}(t)\|$ is constant. Consequently $\|\boldsymbol{r}(t)\|$ is bounded.

2) $\epsilon^{(2)}(t)\equiv 0$*:* The situation here is marginally more complicated. We have from (83) that

$$\Delta \boldsymbol{r}(t) = -K\{\boldsymbol{r}'(t)\boldsymbol{x}(t)\}\boldsymbol{x}(t) - K\{\boldsymbol{r}'(t)\boldsymbol{x}(t)\}\epsilon^{(1)}(t). \tag{108}$$

Hence

$$\Delta \boldsymbol{rs}(t) = -K\{\boldsymbol{rs}'(t)\boldsymbol{x}(t)\}\boldsymbol{x}(t) - K\{\boldsymbol{rs}'(t)\boldsymbol{x}(t)\}\boldsymbol{w}(t) \tag{109}$$

$$\Delta \boldsymbol{rp}(t) = -K\{\boldsymbol{rs}'(t)\boldsymbol{x}(t)\}\boldsymbol{v}(t), \tag{110}$$

where, as before, $\boldsymbol{w}(t)$ and $\boldsymbol{v}(t)$ are respectively the projections of $\epsilon^{(1)}(t)$ onto S and P, respectively.

Equation (109) is simpler than (101); with $\boldsymbol{u}(t)\equiv 0$ in (101) we obtain (109). The following result therefore follows from (103) and (104):

$$\text{if } aKW + e^{-b} < 1, \text{ then } \boldsymbol{rs}(t)\to 0 \text{ as } t\to\infty. \tag{111}$$

Assuming that the above condition for convergence holds, we also have from the Lemma (see (91)),

$$\|\boldsymbol{rs}(N\Delta t)\| \leqslant \|\boldsymbol{rs}(0)\| a(1+KW)(aKW+e^{-b})^{N-1}, \quad \text{for all } N \geqslant 2. \tag{112}$$

The condition for exponential convergence in (111) is the same as the condition for boundedness in (99) except that W occurs in the former in lieu of E_1. The conclusions of the discussion following (99) concerning the requirements on E_1 and K for the boundedness condition to hold are thus applicable here.

Assuming that this condition is satisfied we have, from (110),

$$\begin{aligned}\|\boldsymbol{rp}(N\Delta t)\| &= \|\boldsymbol{rp}(0) - K\sum_{j=0}^{(N-1)\Delta t}\{\boldsymbol{rs}'(j)\boldsymbol{x}(j)\}\boldsymbol{v}(j)\| \\ &\leqslant \|\boldsymbol{rp}(0)\| + KE_1\sum_{j=0}^{\infty}\|\boldsymbol{rs}(j)\| \\ &\leqslant \|\boldsymbol{rp}(0)\| + \frac{a(1+KW)\|\boldsymbol{rs}(0)\|}{1-aKW-e^{-b}} + KE_1\|\boldsymbol{rs}(0)\|.\end{aligned} \tag{113}$$

From (112) and (113) we may thus conclude that in this case $\|\boldsymbol{rs}(N\Delta t)\|\to 0$, $\|\boldsymbol{rp}(N\Delta t)\|$ is bounded and, consequently, $\|\boldsymbol{r}(N\Delta t)\|$ is bounded for all N.

Acknowledgment

We are grateful to D. L. Duttweiler for demonstrating his implementation of an echo canceller and for a stimulating discussion in which he conveyed to us his observations regarding its behavior.

Appendix
The Mixing Condition Is Necessary for Exponential Convergence

The proof of the neccesity of the mixing condition on $\{\boldsymbol{x}(t)\}$ for exponential convergence to zero of the solutions $\boldsymbol{r}(t)$ of (6) is essentially that given by Morgan and Narendra [2]. The main difference is that they dealt with continuous time, while we deal with discrete time. We prove that the existence of an exponentially decreasing bound on $\|\boldsymbol{r}(t)\|$ implies that $\boldsymbol{x}(t)$ is mixing. We do this in two steps—showing that of the following three statements, 1⇒2 and 2⇒3 (we note that 3⇒1 by virtue of our upper bound).

1) There exist positive numbers a and b such that

$$\|\boldsymbol{r}(t_0+N\Delta t)\| \leqslant \|\boldsymbol{r}(t_0)\| ae^{-bN} \text{ for all } t_0 \text{ and } N \geqslant 1. \tag{A1}$$

2) For any unit vector y in R^n there are numbers $N>0$ and $\epsilon>0$, and there is a conical neighborhood[5] (see Fig. 6) C_y of y such that for any $m\geqslant N$ and any t_0

$$\frac{1}{m}\sum_{t\in A(t_0,m,C_w)} K\|\boldsymbol{x}(t)\|^2 \geqslant \epsilon \tag{A2}$$

where[6] $A(t_0,m,C_y)=\{t \mid t=t_0+j\Delta t,\ j=0,1,\cdots,m-1;$ and $x^{\perp}(t)\cap C_y=0\}$. The set $A(t_0,m,C_y)$ is the set of all times, spaced Δt apart in the interval $[t_0,t_0+(m-1)\Delta t]$, where $\boldsymbol{x}(t)$ is not essentially perpendicular to y. Note that if $C_y'\subset C_y$ then $A(t_0,m,C_y)\subset A(t_0,m,C_y')$.

3) $\boldsymbol{x}(t)$ is mixing (see (10) in Section I-E); that is, there exist numbers $T>0$ and $\alpha>0$ such that for any unit vector w and any t_0,

$$\frac{1}{T}\sum_{t=t_0}^{t_0+(T-1)\Delta t}\{w'\boldsymbol{x}(t)\}^2 \geqslant \alpha. \tag{A3}$$

Condition 2 says that on the average $\boldsymbol{x}(t)$ has components of at least a certain size in any given direction; that is, $\boldsymbol{x}(t)$ is not essentially perpendicular to any given direction almost all the time. But this is exactly what the mixing condition says. Thus it seems plausible that 2⇒3; we now see that this is so.

Suppose that condition 2 is satisfied. Then to each unit vector y in R^n there is an associated conical neighborhood C_y. (Since $A(t_0,m,C_y)\subset A(t_0,m,C_y')$ whenever $C_y'\subset C_y$, we may choose any smaller neighborhood than our original neighborhood C_y, and we may wish to do this later.) The conical neighborhoods C_y cover the unit sphere; take a finite subcover $C_{y_1},\cdots,C_{y_s}$. Pick any w on the unit sphere, so $\|w\|=1$. Then there is a y_i such that

[5]A conical neighborhood of y is a set of all points in R^n which can be represented as λu, where $\lambda\in R$ and u is a member of a connected open subset of the unit sphere in R^n which contains y.

[6]We are denoting by $x^{\perp}$ the $(n-1)$ dimensional subspace orthogonal to x.

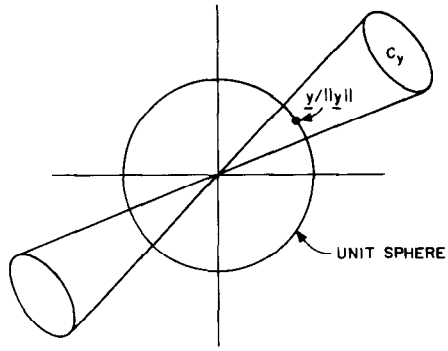

Fig. 6. Conical neighborhood of y.

$w \in C_{y_i}$, and there is an $\epsilon_1 > 0$ such that, for any t_0 and m,

$$\{x'(t)w\}^2 \geq \epsilon_1 \|x(t)\|^2 \tag{A4}$$

for all $t \in A(t_0, m, C_{y_i})$.

Let N and ϵ be those numbers associated with y_i in condition 2. Choosing any $m \geq N$ and any t_0 we see

$$\begin{aligned}\frac{1}{m}\sum_{t=t_0}^{t_0+(m-1)\Delta t}\{x'(t)w\}^2 &\geq \frac{1}{m}\sum_{t\in A(t_0,m,C_y)} K\|x(t)\|^2 \\ &\geq \frac{\epsilon_1}{K}\cdot\frac{1}{m}\sum_{t\in A(t_0,m,C_y)} K\|x(t)\|^2 \\ &\geq \frac{\epsilon\epsilon_1}{K}.\end{aligned} \tag{A5}$$

The second inequality follows from (A4) and the third from condition 2.

Now set $T = \max_{y_i} N(y_i)$ (the y_i make up the finite open cover, and $N(y_i)$ represents the N in condition 2 associated with each y_i), and set

$$\alpha = \frac{1}{K}\inf_{\substack{w\in R^n \\ \|w\|=1}} \epsilon\epsilon_1, \tag{A6}$$

where ϵ and ϵ_1 are associated with each unit vector as in the previous paragraph. It is easy to see that $\alpha > 0$ and $T < \infty$; thus the mixing condition is satisfied.

It remains to show that 1⇒2. We prove this by contradiction. Suppose condition 1 is satisfied, but condition 2 is not. The latter part of the hypothesis means there exists a unit vector $w \in R^n$ so that for any $N > 0$, $\epsilon > 0$ and any conical neighborhood C_w of w, there exist t_0 and t_1 with $t_1 \geq t_0 + (N-1)\Delta t$ and

$$\frac{1}{N}\sum_{t\in A(t_0,N,C_w)} K\|x(t)\|^2 < \epsilon. \tag{A7}$$

Pick N such that $ae^{-bN} \leq 1/2$, where a and b are the constants which appear in condition 1, and pick $\epsilon = 1/16$. Define t_1 to be $t_0 + (N-1)\Delta t$. Define $v(t)$ to be the projection of $x(t)$ on $w^{\perp}$, i.e., $v(t) = [I - ww']x(t)$. The equation $\Delta r(t) = -Kv(t)v'(t)r(t)$ has the stationary solution $r(t) \equiv w$. We show that the equation $\Delta r(t) = -Kx(t)x'(t)r(t)$ has nearly stationary solutions.

Call

$$A(t) = -Kv(t)v'(t) \quad B(t) = -Kx(t)x'(t) - A(t). \tag{A8}$$

Let $Y(t, t_0)$ be the fundamental solution matrix associated with the recursion $\Delta r(t) = A(r)r(t)$, (see (66) in Section IV-A) and consider

$$\Delta r(t) = -Kx(t)x'(t)r(t) = [A(t) + B(t)]r(t), \qquad r(t_0) = w. \tag{A9}$$

Then

$$r(t_1) = w + \sum_{j=t_0+\Delta t}^{t_1} Y(t_1, j)B(j-\Delta t)r(j-\Delta t). \tag{A10}$$

Now

$$\left\|\sum_{j=t_0+\Delta t}^{t_1} Y(t_1,j)B(j-\Delta t)r(j-\Delta t)\right\| \leq \sum_{j=t_0}^{t_1-\Delta t}\|B(j)\| \tag{A11}$$

since, by virtue of the nonincreasing property of $\|r(t)\|$ stated in (8),

$$\|Y(t,j)\| \leq 1 \qquad \|r(j)\| \leq \|r(t_0)\| = \|w\| = 1. \tag{A12}$$

Also

$$\begin{aligned}\|B(j)\| &= \|-Kx(j)x'(j) - A(j)\| \\ &\leq K\|x(j)x'(j)\| + K\|v(j)v'(j)\| \\ &\leq 2K\|x(j)\|^2.\end{aligned} \tag{A13}$$

Thus we have for the right side of (A11),

$$\begin{aligned}\sum_{j=t_0}^{t_1-\Delta t}\|B(j)\| &= \sum_{j\in A(t_0,N,C_w)}\|B(j)\| + \sum_{\substack{j\in[t_0,t_1-\Delta t] \\ j\notin A(t_0,N,C_w)}}\|B(j)\| \\ &\leq 2\sum_{j\in A(t_0,N,C_w)} K\|x(j)\|^2 + \sum_{\substack{j\in[t_0,t_1-\Delta t] \\ j\notin A(t_0,N,C_w)}}\|B(j)\|.\end{aligned} \tag{A14}$$

We may choose C_w so that $\|B(j)\| \leq 1/8N$ for $j \notin A(t_0, N, C_w)$; this means that any two vectors e and f in C_w satisfy either $\|e+f\| \leq 1/8N$ or $\|e-f\| \leq 1/8N$. With such a choice we have

$$\begin{aligned}\sum_{j=t_0}^{t_1-\Delta t}\|B(j)\| &\leq 2\sum_{j\in A(t_0,N,C_w)} K\|x(j)\|^2 + \sum_{\substack{j\in[t_0,t_1-\Delta t] \\ j\notin A(t_0,N,C_w)}}\frac{1}{8N} \\ &\leq 2\frac{1}{16} + N\frac{1}{8N} = \frac{1}{4}\end{aligned} \tag{A15}$$

From (A10) and (A11)

$$\begin{aligned}\|r(t_1)\| &\geq \|w\| - \left\|\sum_{j=t_0+\Delta t}^{t_1} Y(t_1,j)B(j-\Delta t)r(j-\Delta t)\right\| \\ &\geq \|w\| - \sum_{j=t_0}^{t_1-\Delta t}\|B(j)\| \\ &\geq 1 - \frac{1}{4} = \frac{3}{4}.\end{aligned} \tag{A16}$$

But we chose N so that $\|r(t_1)\| = \|r(t_0 + N\Delta t)\| \leq 1/2$. Thus we have our contradiction.

References

[1] M. M. Sondhi and D. Mitra, "New results on the performance of a well-known class of adaptive filters," *Proc. IEEE*, vol. 64, no. 11, pp. 1583–1597, 1976.

[2] A. P. Morgan and K. S. Narendra, "On the uniform asymptotic stability of certain linear nonautonomous differential equations," *SIAM J. Contr.*, vol. 15, no. 1, pp. 5–24, 1977.

[3] P. Eykhoff, *System Identification*. New York: Wiley, 1974, Ch. 7, Ch. 9, Sec. 5.3.

[4] B. Widrow and M. E. Hoff, Jr., "Adaptive switching circuits," in *IRE Wescon Conv. Rec.*, pt. 4, pp. 96–104, 1960.

[5] P. Whitaker, "The MIT Adaptive Autopilot," in *Proc. Self-Adaptive Contr. Symp.*, Wright Air Dev. Center, Wright–Patterson AFB, Ohio, 1959.

[6] K. S. Narendra and L. E. McBride, "Multiparameter self-optimizing systems using correlation techniques," *IEEE Trans. Automat. Contr.*, vol. AC-9, pp. 31–38, 1964.

[7] B. Widrow, P. E. Mantey, L. J. Griffiths, and B. B. Good, "Adaptive Antenna Systems," *Proc. IEEE*, vol. 55, pp. 2143–2159, 1967.

[8] R. W. Lucky, "Automatic equalization for digital communication," *Bell Syst. Tech. J.*, vol. 44, pp. 547–588, 1965.

[9] A. Gersho, "Adaptive equalization of highly dispersive channels for data communication," *Bell Syst. Tech. J.*, vol. 48, pp. 55–70, 1969.

[10] R. W. Lucky, J. Salz, and E. J. Weldon, Jr., *Principles of Data Communication*. New York: McGraw Hill, 1968, ch. 6. See also, R. W. Lucky and H. R. Rudin, "An automatic equalizer for general-purpose communication channels," *Bell Syst. Tech. J.*, vol. 46, pp. 2179–2208, 1967.

[11] M. M. Sondhi, "Closed loop adaptive echo canceller using generalized filter networks," U. S. Patent 3,499,999, Mar. 1970.

[12] J. Kelly and B. F. Logan, Jr., "Self-Adaptive Echo Canceller," U.S. Patent 3,500,000, Mar. 1970.

[13] "Echo Canceller Wins 3,500,000th Patent," *Bell Laboratories Rec.*, p. 126, Apr. 1970.

[14] M. M. Sondhi, "An Adaptive Echo Canceller," *Bell Syst. Tech. J.*, vol. 46, no. 3, pp. 497–511, 1967.

[15] M. M. Sondhi and A. J. Presti, "A self-adaptive echo canceller," *Bell Syst. Tech. J.*, vol. 45, no. 10, pp. 1851–1854, 1966.

[16] F. K. Becker and H. R. Rudin, "Application of automatic transversal filters to the problem of echo supression," *Bell Syst. Tech. J.*, vol. 45, no. 10, pp. 1847–1850, 1966.

[17] S. J. Campanella, H. G. Suyerhoud, and M. Onufry, "Analysis of an adaptive impulse response echo canceller," *COMSAT Tech. Rev.*, vol. 2, no. 1, pp. 1–38, 1972.

[18] N. Demytko and L. K. Mackechnie, "A high speed digital adaptive echo canceller," *Australian Telecomm. Res.*, vol. 7, no. 1, pp. 20–28, 1973.

[19] G. K. Helder and P. C. Lopiparo, "Improving transmission on domestic satellite circuits," *Bell Lab. Rec.*, pp. 202–207, Sept. 1977.

[20] D. L. Duttweiler, "A twelve-channel digital echo canceller," *IEEE Trans. Commmun.*, vol. COM-26, no. 5, pp. 647–653, May 1978.

[21] J. D. Gibson, S. K. Jones, and J. L. Melsa, "Sequentially adaptive prediction and coding of speech signals," *IEEE Trans. Commun.*, vol. COM-22, pp. 1789–1796, 1974.

[22] J. N. Maksym, "Real-time pitch extraction by adaptive prediction of the speech waveform," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 149–153, 1973.

[23] S. Jones, "Adaptive filtering with correlated training samples," Internal Rep., Bell Labs., 1972.

[24] J. K. Kim and L. D. Davisson, "Adaptive linear estimation for stationary M-dependent processes," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 23–31, 1975.

[25] T. P. Daniell, "Adaptive estimation with mutually correlated training sequences," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-6, pp. 12–19, 1970.

[26] G. Ungerboeck, "Theory on the speed of convergence in adaptive equalizers for digital communications," *IBM J. Res. Devel.*, vol. 16, no. 6, pp. 546–555, 1972.

[27] J. R. Rosenberger and E. J. Thomas, "Performance of an adaptive echo canceller operating in a noisy, linear, time-invariant environment," *Bell Syst. Tech. J.*, vol. 50, no. 3, pp. 785–813, 1971.

[28] D. P. Derevitskii and A. L. Fradkov, "Two models for analyzing the dynamics of adaptation algorithms," *Automatika i Tele.*, no. 1, pp. 67–75 (translated), 1974.

[29] N. N. Bogoliubov and J. A. Mitropolski, *Asymptotic Methods in the Theory of Nonlinear Oscillations*. New York: Gordon and Breach, 1961.

[30] R. Z. Khasminskii, "On stochastic processes defined by differential equations with a small parameter," *Theory Prob. Appl.* (USSR), vol. XI, no. 2, pp. 211–228, 1966.

[31] K. Ogata, *State Space Analysis of Control Systems*. Englewood Cliffs, N.J.: Prentice-Hall, 1967, Sec. 6–7.

[32] D. Mitra and M. M. Sondhi, "Summary of results on an adaptive filter using non-ideal multipliers," in 1976 Nat. Telecomm. Conf., Conf. Rec., vol. 1, pp. 8.5-1–8.5-6, Dallas, TX, 1976. See also *IEEE Trans. Automat. Contr.*, vol. AC-24, no. 2, pp. 276–283, 1979.

[33] W. A. Coppel, *Stability and Asymptotic Behavior of Differential Equations*. Boston: D. C. Heath, 1965, Ch. 1, Sec. 3.