and deduces the normal equation by a method of derivation, which introduces some difficulties in the complex case. But even in the real case, it is the wrong way to obtain the results.

It is clear that P(1) is the development of

$$J(\boldsymbol{w}) = E\left\{|y - \boldsymbol{w}^{\dagger}\boldsymbol{x}|^2\right\}$$

where y is the desired signal and $\boldsymbol{x}$ the vector observation, which are supposed to be complex second-order random variables (RV). It is much better to start from (1) than from P(1).

In fact (1) shows that $J(\boldsymbol{w})$ is a distance in the Hilbert space H of second-order RV's. In this space the scalar product between two vectors u and v simply $E(uv^*)$, where the star denotes the complex conjugate. Let us call H_x the subspace of H containing all the RV's on the form $\boldsymbol{h}^{\dagger}\boldsymbol{x}$. The problem is to find the element of H_x giving the minimum distance to y. The solution is well known, and is given by the projection theorem, or

$$\boldsymbol{w}^{\dagger}\boldsymbol{x} = \text{Proj}\left[y|H_x\right] \tag{2}$$

which is the geometrical definition of $\boldsymbol{w}$. To find explicity the solution, it is sufficient to apply the orthogonality principle, or

$$y - \boldsymbol{w}^{\dagger}\boldsymbol{x} \perp\!\!\!\perp H_x \tag{3}$$

which can be written

$$E\left\{\boldsymbol{h}^{\dagger}\boldsymbol{x}(y - \boldsymbol{w}^{\dagger}\boldsymbol{x})^*\right\} = 0, \qquad \forall\, \boldsymbol{h}. \tag{4}$$

Developing this expression we obtain

$$\boldsymbol{h}^{\dagger}(\boldsymbol{p} - \boldsymbol{R}\boldsymbol{w}) = 0 \tag{5}$$

and as this must be valid for all $\boldsymbol{h}$, we obtain $\boldsymbol{R}\boldsymbol{w} = \boldsymbol{p}$, or the normal equation P(2).

This proof is absolutely general and does not introduce any difference between the complex and the real case. It guarantees the fact that it is a minimum, while this point needs a special argument when using differentiation giving only an extremum.

An objection can be made that it makes use of the projection theorem which can be unknown. To overcome this, it is sufficient to prove that any other solution gives a greater mean-square error, which is an indirect proof of the projection theorem, for those not familiar with it.

For this we start from P(1) and we calculate $J(\boldsymbol{R}^{-1}\boldsymbol{p} + \boldsymbol{v})$. By simple algebra this gives

$$J(\boldsymbol{R}^{-1}\boldsymbol{p} + \boldsymbol{v}) = J(\boldsymbol{R}^{-1}\boldsymbol{p}) + \boldsymbol{v}^{\dagger}\boldsymbol{R}\,\boldsymbol{v}. \tag{6}$$

As $\boldsymbol{R}$ is a nonnegative definitive matrix, $\boldsymbol{v}^{\dagger}\boldsymbol{R}\,\boldsymbol{v} \geq 0$, and the result is

$$\forall\, \boldsymbol{v},\ J(\boldsymbol{R}^{-1}\boldsymbol{p} + \boldsymbol{v}) \geq J(\boldsymbol{R}^{-1}\boldsymbol{p}) \tag{7}$$

which completes the proof, or proves the well-known result that the projection gives the minimum distance.

This proof is not only the shortest possible, but also puts the problem in its true framework, which is the concept of Hilbert space of RV's, already known, [1, p.96], [2, p.40], [3, p. 25]. Furthermore, it is the basis of a geometrical interpretation of all the mean-square estimation problems whose extension to constrained problems [5] and nonlinear problems [5] are solved by the same methods.

References

[1] H. Cramer and M. R. Leadbetter, *Stationary and Related Stochastic Processes*. New York: Wiley, 1967.

[2] B. Picinbono, "Signaux déterministes et aléatoires, analyse et modélisation," in *Traitement du signal*, Les Houches session 45. North Holland, 1987.

[3] A. A. Giordano and F. M. Hsu, *Least Square Estimation with Application to Digital Signal Processing*. New York: Wiley, 1985.

[4] B. Picinbono and M. Bouvet, "Constrained Wiener filtering," *IEEE Trans. Inform. Theory*, vol. 33, pp. 160–166, Jan. 1987.

[5] B. Picinbono and Y. Gu, "Mean-square estimation and projections," *Signal Processing*, vol. 19, pp. 1–8, Jan. 1990.

Variable Length Stochastic Gradient Algorithm

Zeev Pritzker and Arie Feuer

Abstract—**This correspondence describes the transversal variable length stochastic gradient (VLSG) algorithm. The algorithm is derived from the stochastic gradient (SG) algorithm which is modified in order to allow dynamic allocation of coefficients of an adaptive filter. The order of the filter and the adaptation step size are changed automatically when an appropriate level of performance is reached during the course of the adaptation process. This way the algorithm results in both fast convergence, typical to low order filters, and good steady state performance, typical of high order filters.**

I. Introduction

The stochastic gradient (SG) is a very common adaptive algorithm [1]–[3], [6]. Probably the most common SG algorithm is the least mean square (LMS). It is a very simple algorithm and that is the main reason for its popularity. However, a major problem with the LMS is its slow convergence compared to other algorithms which are clearly more complicated (e.g., the recursive least squares algorithm). As a result, considerable effort has been directed towards improving the convergence rate of the LMS while preserving its basic simplicity.

In this correspondence we describe a new modification of the LMS and from the results we have, quite a promising one. Our idea is based on the following two observations: Generally, with the LMS, the lower the dimension of the regression vector, the faster the algorithm convergence. On the other hand, the higher the dimension of the regression vector, the better the algorithms steady state performance. To accommodate these two contradictory goals we propose an LMS which can change its dimension—initially low dimension to achieve initial fast convergence, and gradually increased dimension to finally give the desired steady state performance. We called this algorithm the variable length LMS (VL-LMS).

While an analysis of the VL-LMS performance is beyond the scope of this paper we make use of the close relationship between the steepest descent (SD) algorithm and LMS—the LMS can be viewed as an approximation of the SD. We apply the variable length approach to the SD to get what we called the VLSD and analytically show the convergence rate improvement possibilities. This provides the motivation for applying the variable length approach to the LMS and, as a matter of fact, to many other SG algorithms. Extensive simulations verified the improved convergence rate of the VL-LMS over the standard LMS, and a sample of these results is presented here.

II. The Variable Length Steepest Descent (VLSD)

The general problem we address can be presented as follows.

Given a vector $\{X_n\} \in C^N$ and a scalar $\{d_n\}$ stochastic sequences find a vector of weights $W_n \in C^N$ which will minimize the

Manuscript received June 23, 1988; revised May 8, 1990.

The authors are with the Department of Electrical Engineering, Technion Israel Institute of Technology, Haifa 3200, Israel.

IEEE Log Number 9042279.

mean-square error (MSE)

$$J = E\{|e_n|^2\} = E\{|d_n - W^H X_n|^2\}. \quad (1)$$

$E\{\ \}$ denotes the "expected value of $\{\ \}$" and $(\ )^H$ is the complex conjugate transpose. The solution to this problem is well known and given by

$$W_{\text{opt}} = R_N^{-1} p \quad (2)$$

where

$$R_N = E\{X_n X_n^H\} \quad (3)$$

and

$$p = E\{d_n^* X_n\}. \quad (4)$$

W_{opt} can be computed in an iterative way by means of the SD algorithm

$$W_{n+1} = (I - \mu R_N) W_n + \mu p \quad (5)$$

which is known to converge, with properly chosen step size, to W_{opt}. (For a more detailed description and discussion see, e.g., [6].)

We note that in all adaptive filtering applications $X_n = [x_n, x_{n-1}, \cdots, x_{n-N+1}]^T$ where x_n is a stochastic stationary sequence, so that

$$R_N(i, j) = E\{x_{n-i+1} x_{n-j+1}^*\} = r(i - j) \quad (6)$$

where $R_N(i, j)$ is the (i, j)th element of R_N and $r(k)$ is the kth sample of the autocorrelation sequence of x_n.

By substituting (2) into (1) we get the minimal MSE for the N-dimensional case

$$J_{\min}(N) = E\{|d_n|^2\} - p^H R_N^{-1} p \quad (7)$$

while the learning curve (describing the MSE at each iteration) is

$$J_n(N) = J_{\min}(N) + (W_n - W_{\text{opt}})^H R_N (W_n - W_{\text{opt}}). \quad (8)$$

The rate of convergence of J_n depends on the convergence of W_n. W_n has several modes contributing to its convergence and clearly the slowest will dominate its rate of convergence. So we will use the slowest mode of W_n's convergence as our measure of the convergence rate of W_n. It is well known (see, e.g., [6]), that the choice

$$\mu_{\text{opt}}(N) = \frac{2}{\alpha_{\max}(N) + \alpha_{\min}(N)} \quad (9)$$

will result in the fastest convergence of the SD. $\alpha_{\max}(N)$ and $\alpha_{\min}(N)$ are the largest and smallest eigenvalues of R_N, respectively. With this choice of μ the slowest mode is given by

$$\beta_N = |1 - \mu_{\text{opt}} \alpha_{\max}(N)| = |1 - \mu_{\text{opt}} \alpha_{\min}(N)| = \frac{S_\alpha(N) - 1}{S_\alpha(N) + 1} \quad (10)$$

where

$$S_\alpha(N) = \alpha_{\max}(N)/\alpha_{\min}(N) \quad (11)$$

is the eigenvalue spread of R_N.

It can be readily seen by (10) that β_N is a monotonically increasing function of $S_\alpha(N)$.

Next we are going to show that $S_\alpha(N)$ is a monotonically nondecreasing (strictly increasing in most cases) function of the dimension N.

The computation of the eigenvalues of R_N is not an easy task, even for small values of N (see, e.g., [5]) and no closed form expressions exist for these eigenvalues or the corresponding eigenvalue spread. Instead, we bound $S_\alpha(N)$ by $S_\alpha(N-1)$ and achieve our purpose this way.

Since R_N is Hermitian it can be partitioned as follows:

$$R_N = \begin{bmatrix} R_{N-1} & r_{N-1} \\ r_{N-1}^H & r(0) \end{bmatrix} \quad (12)$$

where $r_{N-1} \in C^{N-1}$. Then we have the following:

Theorem 1: Let R_N be a positive definite matrix partitioned as in (12) where R_{N-1} is its principal submatrix. Then

$$S_\alpha(N) \geq S_\alpha(N-1) \quad (13)$$

for all $N > 1$. Furthermore, let an eigenvector of R_{N-1} corresponding to either $\alpha_{max}(N-1)$ or $\alpha_{min}(N-1)$ be not *orthogonal to r_{N-1}. Then*

$$S_\alpha(N) > S_\alpha(N-1) \quad (14)$$

for $N > 1$.

The proof of this theorem follows directly from results in [7] and [8] where the eigenvalues of R_{N-1} and R_N are shown to have an interlacing property.

We have so far shown that β_N is a monotonically increasing function of $S_\alpha(N)$ and that $S_\alpha(N)$ is a monotonically increasing (in most cases) function of N. Thus, clearly, *β_N is a monotonically increasing function of N*. This means that the larger N is, the slower the convergence of W_n to W_{opt}.

We, however, are interested in the behavior of the MSE. Using results from [6] we know that the slowest natural mode of the MSE, which, again dominates its behavior, is given by

$$J_n(N) = J_{\min}(N) + (J_0 - J_{\min}(N))\beta_N^{2n}.$$

To apply our approach we must guarantee that starting with the same initial MSE, after the first iteration the MSE for N is smaller than the MSE for $(N+1)$. For this we need the following:

Assumption A1: The minimum MSE levels $J_{\min}(N)$ and $J_{\min}(N+1)$ satisfy the following inequality for $N > \mathcal{N}$

$$\frac{J_0 - J_{\min}(N)}{J_0 - J_{\min}(N+1)} > \frac{1 - \beta_{N+1}^2}{1 - \beta_N^2}. \quad (15)$$

Assumption A1 together with the monotonicity for β_N will guarantee that

$$J_1(N) < J_1(N+1)$$

and applying the proposed variable length approach will give the desired improved performance.

Note that in practice it is common to have $J_0 \gg J_{\min}(N)$. Then, because of the monotonicity of β_N, assumption A1 is readily satisfied.

The proposed VLSD algorithm is then as follows. The adaptation is started with N weights and the algorithm is switched to $N+1$ weights at an appropriate moment (e.g., any moment before n_0 in Fig. 1). The added weight is initially set to zero. The process is repeated until the desired filter order is reached. The advantage of the VLSD over SD is illustrated in Fig. 2.

The choice of the initial order N should be made with caution in order to satisfy Assumption A1. If N is chosen to be too small $J_{\min}(N)$ may be not far enough from J_0 to satisfy (15). Simulations have shown that in practice choosing a sufficiently large N(3 to 6 in most applications) will solve the problem.

The ideal switching point, n_a, is illustrated in Fig. 1, and an exact expression for it as a function of $J_{\min}(N)$, $J_{\min}(N+1)$, β_N and β_{N+1} can be derived. This expression is quite complicated and in practice the following simple technique has been successfully implemented. A small $\phi > 0$ is set and the algorithm is switched to a higher order when the MSE decays below $J_{\min} + \phi$, i.e, whenever $J_n < J_{\min} + \phi$.

So far we have described the VLSD. This algorithm cannot be of much practical use since exact knowledge of input statistics is required for its implementation. However, it provides valuable in-

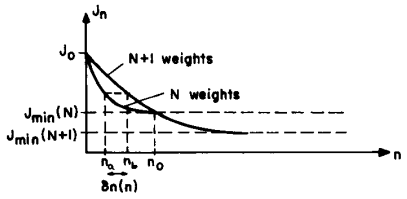

Fig. 1. Learning curves of Nth and $(N+1)$th order filters.

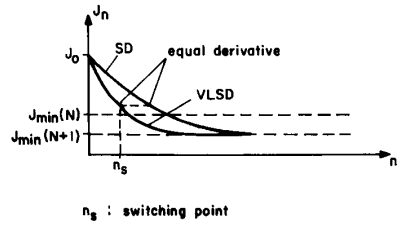

Fig. 2. Learning curves of the $(N+1)$th order SD algorithm and the VLSD algorithm with $N = N$, $\bar{N} = N + 1$.

sights and motivation for the forthcoming VL-LMS algorithm. The results of this section were verified by computer simulations which confirm the validity of our analysis. The simulation results are given in Section IV.

III. The Variable Length LMS Algorithm (VL-LMS)

Taking the SD algorithm in (5) and replacing R_N and p with their instantaneous estimates $X_n X_n^H$ and $d_n^* X_n$, respectively, one gets the well-known LMS algorithm

$$W_{n+1} = W_n + \mu e_n^* X_n \tag{16}$$

where

$$e_n = d_n - W_n^H X_n. \tag{17}$$

The proposed VL-LMS algorithm is then as follows. The recursive algorithm (16) is initially used with N weights. The algorithm is switched to $N+1$ when the MSE is close to $J_{\min}(N)$ where the added weight is initially set to zero. The process is repeated until the desired filter order is reached. In fact the VL-LMS algorithm is the stochastic version of the VLSD algorithm just as much as the LMS is the stochastic version of the SD. Between the switching points the VL-LMS behaves identically to the corresponding fixed length LMS with the same number of weights and same initial conditions. It is known, [6], that the learning curve of the LMS exhibits behavior similar to that of the corresponding (same step size) SD. Hence, the learning curve of the VL-LMS will exhibit behavior similar to that of the VLSD.

The choice of the switching points, however, has to be reconsidered. Recall that in Section II, $J_{\min}(N)$ were assumed to be known for each N. This assumption is not valid in practical use of the VL-LMS. Still, in many applications, some *a priori* knowledge enables the derivation of a set of reasonable estimates for these values. For example, in the adaptive equalization application the algorithms input $\{x_n\}$ is the output from a linear channel excited by an i.i.d data sequence. Since voiceband communication channels are only allowed to vary within certain limits (such as CCITT specifications for allowed amplitude and phase responses) the channels output $\{x_n\}$ belongs to a bounded population of stochastic processes for which a reasonable set of estimates of $J_{\min}(N)$ can be generated *a priori*.

In order to detect switching points we need to estimate recursively the MSE, $J_n(N)$. A simple and robust estimate of $J_n(N)$ can be obtained by performing "exponential smoothing" of the square error in (17). So

$$\hat{J}_n(N) = (1 - \Omega)\hat{J}_{n-1}(N) + |e_n|^2 \tag{18}$$

where $0 < \Omega < 1$.

Another difference between the VLSD and the VL-LMS is the choice of the step size $\mu(N)$. It cannot be chosen equal to that of (9) since the LMS algorithm requires smaller μ in order to converge (see, e.g., [2]). The bound suggested in [2] is

$$\mu_{\max}(N) = 2/[3trR_N] = 2/[3Nr(o)].$$

However, to get a smoother behavior for the LMS we chose the value

$$\mu(N) = 0.2/[Nr(0)] \tag{19}$$

proposed in [6].

It should be noted that this choice for our initial N may well be larger than the allowable for final desired filter length $\bar{N}$.

The "misadjustment" is defined for the fixed length LMS as the ratio of the excess MSE due to gradient noise to the minimal achievable MSE. Its steady state value in a stationary environment was given approximately by [6] as

$$M(N) = 1/[2\mu Nr(o)]. \tag{20}$$

It has been observed that in practice the noisiness of the LMS learning curves remains very much the same during the learning period and the steady state. $M(N)$ can then serve as a measure of gradient noise level during the adaptation process. For the VL-LMS algorithm substitution of (19) into (20) results in

$$M(N) = 1/[2\mu(N)Nr(o)] = 0.1 = \text{constant}. \tag{21}$$

This reflects an important property of the VL-LMS algorithm: Despite the fact that the step size is increased (typically by an order of magnitude, beyond even the stable range of the fixed length LMS) during early stages of adaptation, the level of gradient noise remains constant. If such large step sizes were used in fixed length LMS a very noisy learning curve or even divergence would result. Note that the steady state misadjustment of the VL-LMS is the same as for the fixed length LMS of order $\bar{N}$.

. The VL-LMS algorithm is summarized as follows:

1) < Initialize >
   1.1 Set a small $\phi > 0$
   1.2 Set a small $\Omega > 0$
   1.3 Set an array of empirically estimated MSE levels $J_{\min}(N)$ to $J_{\min}(\bar{N} - 1)$
   1.4 $N = N$
   1.5 $\mu = 0.2/[Nr(0)]$
   1.6 $W_0 = 0$
2) For $n = 0$ to infinity do:
   2.1 $e_n = d_n - W_n^H X_n$
   2.2 $W_{n+1} = W_n + \mu e_n^* X_n$
   2.3 $\hat{J}_n = (1 - \Omega)\hat{J}_{n-1} + |e_n|^2$
   2.4 if $\hat{J}_n > J_{\min}(N) + \phi$ go to 2.1
   2.5 < Increase filter order >
      2.5.1 if $N = \bar{N}$ go to 2.1
      2.5.2 $N = N + 1$
      2.5.3 $W_n(N) = 0$ < initialize at zero the added weight >
      2.5.4 $\mu = 0.2/[Nr(o)]$
      2.5.5 go to 2.1.

IV. Computer Simulations

In order to test the proposed algorithm extensive simulations have been performed. The inverse modeling application (see [6]) was chosen in order to demonstrate the VLSD and the VL-LMS algo-

rithms performances. The variable parameters of the simulation were the plant's impulse response and the minimal and final filter lengths, N and $\bar{N}$, respectively. The performance of the VLSD and the VL-LMS were compared with the performance of the conventional SD and LMS algorithms applied to an adaptive FIR filter of fixed length $\bar{N}$.

R_N, p, and $J_{\min}(N)$, $(N \le N < \bar{N})$ were computed from the plant's impulse response and the input noise variance. The SD and the LMS algorithms were run using (5) and (16), respectively, with identical step size given by (19). The VL-LMS algorithm was implemented as described at the end of Section III. The VLSD algorithm was implemented as described in Section II, except the choice of $\mu(N)$ and the switching moments, which were chosen to be identical to those of the VL-LMS.

Fig. 3 shows the results of the runs of the four algorithms on the same scale. The plant impulse response is $(0.5)^n$ for $n = 0, 1, 2, 3$, and $N = 2$, $\bar{N} = 8$. The curves for LMS and VL-LMS are the averages of 10 independent runs. As expected, the LMS and the VL-LMS learning curves closely follow those of the SD and VLSD. Clearly, the variable length algorithms converge considerably faster than their fixed length counterparts. Despite the fact that the step size which is initially used for the VL-LMS is 4 times that of the LMS ((19) for $N = 2$ and $\bar{N} = 8$), the VL-LMS learning curve exhibits the same low level of gradient noise as that of the LMS. These results are in full agreement with our analysis.

Fig. 4 corresponds to plant with impulse response $(0.5)^n$, $n = 0, 1$. Again $N = 2$, $\bar{N} = 8$ and we observe a clear advantage of the VL-LMS over the LMS.

The choice of the switching parameter ϕ was made empirically. The best value of ϕ was found to be 0.15 with initial MSE normalized to unity. The sensitivity of the algorithm to ϕ appears to be very low: The behavior of the VL-LMS algorithm was practically unaffected when ϕ was changed in the 0.06–0.2 range. Clearly, this reflects low sensitivity of the algorithm to the estimated $J_{\min}(N)$.

We have also applied the VL approach to another SG algorithm used for blind equalization (see [4]). The key idea is to avoid the use of training sequences and then the MSE turns out to not be an appropriate performance criterion [1], [3]. The performance surface is nonquadratic in this case, and a good measure of a blind equalizer's performance (residual intersymbol interference in a digital communication system) is

$$D = \frac{\sum_{k \neq 0} V_k^2}{V_0^2}$$

where $\{V_k\}$ is the impulse response of the overall communication system (the channel in series with the equalizer) sampled at the symbol transmission rate. For more details see, e.g., [1]. Zero distortion ($D = 0$) means perfect equalization. The condition $D = -15$ dB is commonly referred to as "open eye." When the blind equalizer reaches the open eye condition, the equalization can be switched to a much faster "decision directed" algorithm [3]. The time required for convergence to the -15 dB level is therefore of paramount importance.

In Fig. 5 we show the evolution of the distortion for the fixed length blind equalizer with 30 complex weights (FLBE-30) which is typical for practical implementations. An alternative approach proposed in [4] is to use FLBE-15 and to pad zeroed weights to get the 30 taps before switching to the decision directed mode. In Fig. 5 we show both the FLBE-15 and the corresponding VLBE. We clearly observe the faster convergence of VLBE to the open eye level. Using the variable length filter strategy reduces the overall convergence to that of the conventional equalizer without the penalty of training sequence.

V. Conclusion

In this correspondence a new transversal variable length stochastic gradient (VLSG) algorithm was proposed. The key idea is to adjust the number of the adaptive filter's weights (filter length) dynamically. This way we accomplish both fast initial convergence, typical to small number of filter weights and low steady state MSE, typical to large number of filter weights.

We have concentrated here on the VL-LMS but the same approach can be implemented on other SG algorithms as demonstrated in our simulations. Also, we believe that additional performance improvement can be gained by clever design of a variable length stochastic gradient algorithm in the nonstationary input case. This, however, is a subject for further study.

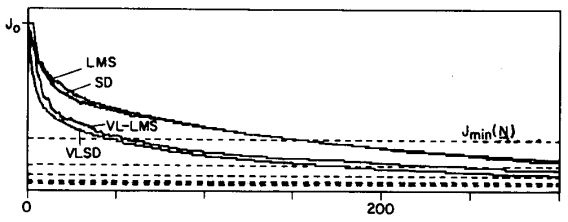

Fig. 3. Learning curves of the SD and the LMS with $\bar{N} = 8$ and the VLSD, VL-LMS with $N = 2$, $\bar{N} = 8$.

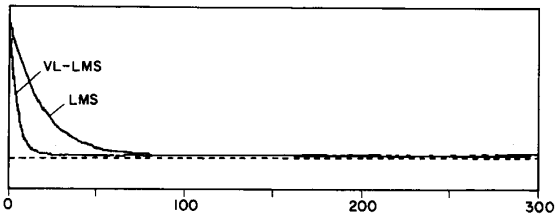

Fig. 4. A comparison between LMS and VL-LMS performances ($N = 2$, $\bar{N} = 8$ plant length is 2).

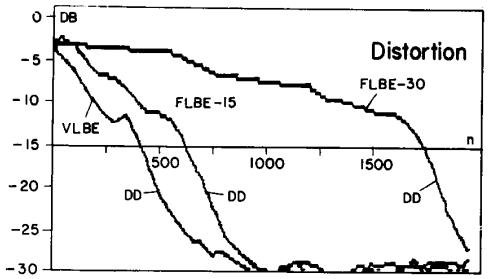

Fig. 5. A comparison of variable length blind equalizer (VLBE) and fixed length blind equalizer (FLBE) performances.

References

[1] A. Benveniste and M. Rouget, "Blind equalizers," *IEEE Trans. Commun.*, vol. COM-32, no. 8, pp. 871–883, Aug. 1984.

[2] A. Feuer and E. Weinstein, "Convergence analysis of LMS filters with uncorrelated Gaussian data," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 222–230, Feb. 1985.

[3] D. Godard, "Self-recovering equalization and carrier tracking in two-dimensional data communication systems," *IEEE Trans. Commun.*, vol. COM-28, no. 11, pp. 1867–1875, Nov. 1980.

[4] Z. Pritzker, "The VLSG algorithm and its application to blind equalization in digital communication systems," M.Sc. thesis, Technion, Israel Institute of Technology, 1988.

[5] A. Ralston and P. Rabinovitz, *A First Course in Numerical Analysis*. New York: McGraw-Hill, 1986.

[6] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1985.

[7] A. Dembo, "Bounds on the extreme eigenvalues of positive-definite Toeplitz matrices," *IEEE Trans. Inf. Theory*, vol. 34, no. 2, pp. 352-355, Mar. 1988.

[8] D. Slepian and H. J. Landau, "A note on the eigenvalues of Hermitian matrices," *SIAM J. Math. Anal.*, vol. 9, no. 2, pp. 291-297, Apr. 1978.

Cumulant Series Expansion of Hybrid Nonlinear Moments of Complex Random Variables

Gaetano Scarano

***Abstract*—In this correspondence a general theorem for zero-memory nonlinear (ZNL) transformations of complex stochastic processes is presented. It will be shown that, under general conditions, the cross covariance between a stochastic process and a distorted version of another process can be represented by a series of cumulants. The coefficients of this cumulant expansion are expressed by the expected values of the partial derivatives, appropriately defined, of the function describing the nonlinearity.**

The theorem includes as a particular case the well-known invariance property (Bussgang's theorem) of Gaussian processes, while holding for any joint distribution of the processes. The expansion in cumulants constitutes an effective means of analysis for higher order moment based estimation procedures involving non-Gaussian complex processes.

I. Introduction

Bussgang's theorem [1] (as extended to the complex case in [2]) states that the cross correlation between two jointly normal, zero-mean, stationary complex stochastic processes $x(t)$ and $y(t)$ is proportional to the cross correlation between $x(t)$ and $z(t)$, a (complex) zero-memory nonlinear (ZNL) transformation of $y(t)$

$$R_{XZ}(\tau) = E\{x(t) \cdot \bar{z}(t-\tau)\} = K \cdot R_{XY}(\tau)$$
$$= K \cdot E\{x(t) \cdot \bar{y}(t-\tau)\}$$

in which $z(t) = g[y(t)]$. An overbar denotes complex conjugation. The proportionality factor

$$K = \frac{1}{\sigma_y^2} E\{y(t) \cdot \bar{g}[y(t)]\}$$

was derived in [2], in which $\sigma_y^2 = E\{|y|^2\}$ is the variance of $y(t)$. This is also referred to as the invariance property of complex Gaussian processes.

In this correspondence, it is shown that the invariance property is a special case for Gaussian processes of a more general theorem that holds for any distribution of the processes. In fact, it will be demonstrated that, when the function $g(.)$ that can be expressed in the form $g(x) = f(u, w)|_{\substack{u=x \\ w=\bar{x}}}$ where $f(u, w)$ is analytic both in u and w, the cross covariance $E\{x(t) \cdot \bar{z}(t-\tau)\}$ can be expanded in a series of cumulants weighted by the expected values of the partial derivatives of the function $f(u, w)$. This expansion in cumulants constitutes a useful means of analysis for higher order moment based estimation procedures involving complex, non-Gaussian processes.

Manuscript received September 10, 1990.

The author is with C.N.R., Istituto di Acustica "O. M. Corbino," I-00189 Rome, Italy.

IEEE Log Number 9042268.

II. The Cumulant Expansion

The class of analytic conjugate functions in a field A is introduced and defined as follows:

$$\mathcal{C}_A^* = \left\{g(x): g(x) = f(u, w)\big|_{\substack{u=x \\ w=\bar{x}}}, \right.$$
$$\left. f(u, w) \text{ analytic in } (A, A)\right\}.$$

For analytic conjugate functions, the differentiation is defined in terms of differentiation of the analytic function $f(u, w)$, i.e.,

$$\frac{\partial^{p+q}}{\partial x^p \partial \bar{x}^q} g(x) = g^{(p,q)}(x)$$
$$= \frac{\partial^{p+q}}{\partial u^p \partial w^q} f(u, w)\Big|_{\substack{u=x \\ w=\bar{x}}}$$
$$= f^{(p,q)}(x, \bar{x}).$$

The expected value is denoted by $E\{.\}$, complex conjugation by the overbar, and the random variables (RV's) extracted at the instants t and $t - \tau$ from processes $x(t)$ and $y(t)$ by X and Y, respectively.

In order to simplify the notations in the following, the case $X = Y$ is considered first. The extension to the bivariate case is straightforward and does not affect the essence of the development that follows.

Theorem 1: Let $g(.)$ be a conjugate analytic ZNL transformation in a field A, on which is defined a circularly symmetric complex random variable $(CRV)X$[1]. Then

$$E\{x \cdot \bar{g}(x)\} = \sum_{q=0}^{\infty} \frac{1}{(q+1)!q!} \bar{E}\{g^{(q+1,q)}(x)\} \cdot k_X^{(q+1,q+1)}$$

where $k_X^{(q+1,q+1)}$ is the complex cumulant of order $(q+1, q+1)$ of the bidimensional $CRV(X, \bar{X})$.

Proof: Let

$p_X(x) = p_{X_r X_i}(x_r, x_i)$	be the joint probability density function (p.d.f.) of the real and the imaginary part of $X = X_r + jX_i$;
$P_X(s, v) = E\{e^{\xi x_r + \eta x_i}\}$ $= E\{e^{sx + v\bar{x}}\}$ $(s = \frac{1}{2}(\xi - j\eta),$ $v = \frac{1}{2}(\xi + j\eta))$	the (complex) moment generating function (m.g.f.) of $(X, \bar{X})$;
$C_X(s, v) = \log P_X(s, v)$	the cumulant generating function (c.g.f.) of $(X, \bar{X})$.

Differentiating with respect to s both sides of

$$C_X(s, v) = \log P_X(s, v)$$

yields

$$P_X(s, v) \cdot C_X^{(1,0)}(s, v) = P_X^{(1,0)}(s, v)$$

where, generically, the subscript $^{(p,q)}$ denotes partial differentiation p times with respect to s and q times with respect to v.

Again, differentiating r times with respect to s and t times with respect to v, the Leibnitz theorem for functions of two variables is obtained:

$$P_X^{(r+1,t)}(s, v) = \sum_{l=0}^{r} \binom{r}{l} \sum_{n=0}^{t} \binom{t}{n} P_X^{(l,n)}(s, v)$$
$$\cdot C_X^{(r+1-l,t-n)}(s, v). \quad (1)$$

[1]For simplicity, moments and cumulants are taken around the origin, which is supposed to be enclosed in the field A. More generally, the theorem holds replacing x by $x - x_0$ (for $x_0 \in A$), and considering moments and cumulants around x_0.

1053-587X/91/0400-1001$01.00 © 1991 IEEE