Proc. Eusipco Signal Processing, pp. 1405-1409, Trieste, Italy, Sep. 1996.

Robustness and Convergence of Adaptive Schemes in Blind Equalization and Neural Network Training

ALI H. SAYED and MARKUS RUPP

Department of Electrical and Computer Engineering
University of California
Santa Barbara, CA 93106–9560.

Abstract— **We pursue a time-domain feedback analysis of adaptive schemes with nonlinear update relations. We consider commonly used algorithms in blind equalization and neural network training and study their performance in a purely deterministic framework. The derivation employs insights from system theory and feedback analysis, and it clarifies the combined effects of the step-size parameters and the nature of the nonlinear functionals on the convergence and robustness performance of the adaptive schemes.**

I. INTRODUCTION

In recent work [1], [2], the authors have formulated a time-domain feedback approach for the analysis and design of adaptive schemes with emphasis on robust performance and improved convergence in the presence of measurement noise and modeling uncertainties. In particular, we have addressed the following two issues:
1. We have shown how to select the adaptation gain (step-size) in order to guarantee a robust behaviour in the presence of noise and modeling uncertainties.
2. We have also shown how to select the adaptation gain in order to guarantee faster convergence.
In this paper, we briefly outline extensions of this formulation to adaptive schemes that involve nonlinear update laws, with special emphasis on the Perceptron Learning Algorithm (PLA, for short) in neural network training and on blind and non-blind equalization schemes in communications. By so doing, we further highlight some common features that exist between neural network structures and blind equalization structures. However, in blind equalization, some complications arise that require a closer analysis. These complications are primarily due to i) the complex nature of the signals involved, ii) to the constant modulus constraints on the signals in the system, and iii) to the blind mode of operation itself.

<ins>**Notation.**</ins> We use small boldface letters to denote vectors, "$*$" for Hermitian conjugation, "T" for transposition, and $\|\mathbf{x}\|$ for the Euclidean norm of a vector. All vectors are column vectors except for the input data vector denoted by $\mathbf{u}_i$, which is taken to be a row vector. We also use the shift operator q^{-1}, defined by $q^{-1}s(i) = s(i-1)$, to denote the unit time delay.

II. NONLINEAR ADAPTIVE SCHEMES

The Perceptron consists of a linear combiner, whose column weight vector we denote by $\mathbf{w}$, followed by a nonlinearity f, known as an activation function. It is depicted in Figure 1 where $\mathbf{u}$ denotes an input (row) vector. A common choice for f is the sigmoid function $f_\beta(z) = \frac{1}{1+e^{-\beta z}}$, with $\beta > 0$ [3]. But, more generally, it can be any monotonically increasing function.

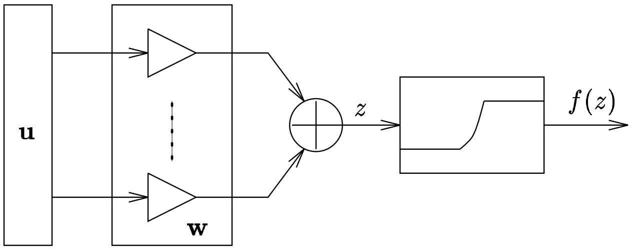


Fig. 1. *The Perceptron structure.*

Let $\{y(i)\}$ be a collection of output (or reference) values that are assumed to belong to the range of the activation function $f(\cdot)$, i.e., there exists a $\mathbf{w}$ and a row input vector $\mathbf{u}_i$ such that $y(i) = f(\mathbf{u}_i\mathbf{w})$. In supervised learning, the Perceptron is presented with given input-output data $\{\mathbf{u}_i, d(i)\}$, where $d(i)$ are possibly noisy or perturbed versions of $y(i)$, say $d(i) = y(i) + v(i)$, and the objective is to estimate $\mathbf{w}$. The PLA computes recursive estimates of $\mathbf{w}$ as follows:

$$\mathbf{w}_i = \mathbf{w}_{i-1} + \mu(i)\mathbf{u}_i^T \left[d(i) - f(\mathbf{u}_i\mathbf{w}_{i-1})\right] \ , \qquad (1)$$

This material was based on work supported in part by the National Science Foundation under Award No. MIP-9409319.

This work was partially performed while M. Rupp was a post-doctoral fellow in the ECE Dept., UCSB, CA. He is currently with the Wireless Technology Research Dept., Lucent Technology, 791 Holmdel-Keyport Rd., Holmdel NJ 07733–0400.

where $\mu(i)$ denotes the step-size parameter (possibly time-variant).

A similar nonlinear training structure arises in channel equalization, as depicted in Fig. 2. The figure shows a sequence $\{s(i)\}$ (usually complex and of constant modulus) being transmitted through an unknown channel $C(q^{-1})$. The receiver is assumed to have an adaptive M-th order FIR structure with weights $\mathbf{w}_{i-1}$, followed by a nonlinear decision device f. The output of the decision device is used to compute an error quantity $e_o(i)$ that is employed in the training algorithm:

$$\mathbf{w}_i = \mathbf{w}_{i-1} + \mu(i)\mathbf{u}_i^* e_o(i) \ , \qquad (2)$$

with $\mathbf{u}_i = \left[\begin{array}{ccc} u(i) & \ldots & u(i-M+1) \end{array} \right]$. The definition of the error quantity $e_o(i)$ depends on whether the equalizer operates in a blind mode or not, which in turn determines the nature of the additional measurement used in Fig. 2. In non-blind operation, the measurement is $s(i-D)$ (a delayed version of $s(i)$) and $e_o(i) = s(i-D) - f(\mathbf{u}_i\mathbf{w}_{i-1})$. In blind operation, $e_o(i)$ is taken as $e_o(i) = f(\mathbf{u}_i\mathbf{w}_{i-1}) - \mathbf{u}_i\mathbf{w}_{i-1}$. We assume for our analysis that there exists an optimal receiver $\mathbf{w}$ with such a structure, FIR followed by the nonlinearity, and which guarantees detection, viz., $f(\mathbf{u}_i\mathbf{w}) = s(i-D)$.

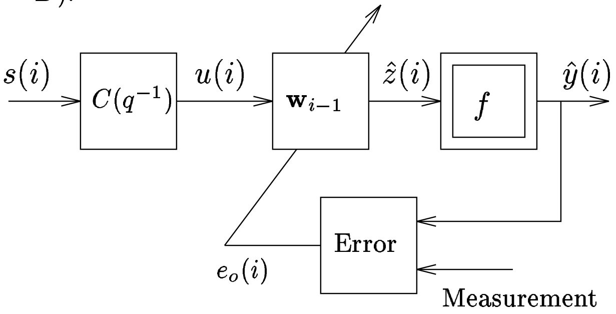


Fig. 2. *Structure of the nonlinear adaptive equalizer.*

Table I lists several nonlinear functions that have been used in channel equalization (see, e.g., [4]):

Equalization type/algorithm	$f[z]$
Direct-decision 2-PSK	sign$[z]$
Direct-decision equalizer	dec$[z]$
CMA (Godard 2-2)	$z\lvert z\rvert^2$
Norm. CMA (Godard 1-2)	$\frac{z}{\lvert z\rvert}$
Sato's algorithm	γsign$[z]$

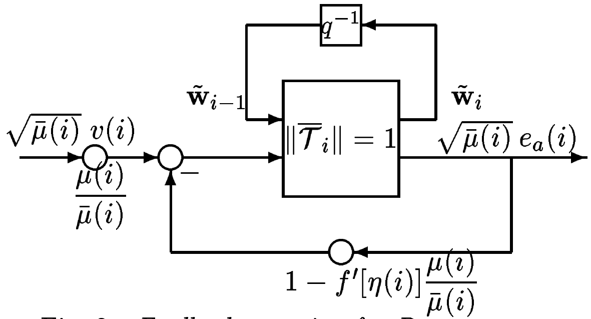
TABLE I

Nonlinear devices for equalization.

III. PERCEPTRON TRAINING

The following quantities are useful for our analysis: $\tilde{\mathbf{w}}_i = \mathbf{w} - \mathbf{w}_i$, $e_a(i) = \mathbf{u}_i\tilde{\mathbf{w}}_{i-1} = z(i) - \hat{z}(i)$, $e_p(i) = \mathbf{u}_i\tilde{\mathbf{w}}_i$, and $\bar{\mu}(i) = 1/\|\mathbf{u}_i\|^2$. It has been shown in [5] that the following equality holds for all possible choices of $\mu(i)$:

$$\frac{\|\tilde{\mathbf{w}}_i\|^2 + \bar{\mu}(i)e_a^2(i)}{\|\tilde{\mathbf{w}}_{i-1}\|^2 + \bar{\mu}(i)e_p^2(i)} = 1 \ , \qquad (3)$$



q^{-1}
$\tilde{\mathbf{w}}_{i-1}$
$\tilde{\mathbf{w}}_i$
$\sqrt{\bar{\mu}(i)}\ v(i)$
$\frac{\mu(i)}{\bar{\mu}(i)}$
$-$
$\|\overline{\mathcal{T}}_i\| = 1$
$\sqrt{\bar{\mu}(i)}\ e_a(i)$
$1 - f'[\eta(i)]\frac{\mu(i)}{\bar{\mu}(i)}$

Fig. 3. *Feedback mapping for Perceptron.*

which establishes the existence of a lossless mapping $\overline{\mathcal{T}}_i$ from the signals $\{\tilde{\mathbf{w}}_{i-1}, \sqrt{\bar{\mu}(i)}e_p(i)\}$ to the signals $\{\tilde{\mathbf{w}}_i, \sqrt{\bar{\mu}(i)}e_a(i)\}$.

If we further apply the mean-value theorem to the activation function $f(z)$, and write

$$f[\mathbf{u}_i\mathbf{w}] - f[\mathbf{u}_i\mathbf{w}_{i-1}] = f'[\eta(i)]e_a(i),$$

for some point $\eta(i)$ along the segment connecting $\mathbf{u}_i\mathbf{w}$ and $\mathbf{u}_i\mathbf{w}_{i-1}$, we can further show that

$$-\bar{\mu}^{\frac{1}{2}}(i)e_p(i) = \frac{\mu(i)}{\bar{\mu}^{\frac{1}{2}}(i)}v(i) - \left[1 - f'[\eta]\frac{\mu(i)}{\bar{\mu}(i)}\right]\bar{\mu}^{\frac{1}{2}}(i)e_a(i) \ .$$

This relation shows that the overall mapping from the *original* (weighted) disturbances $\sqrt{\bar{\mu}(\cdot)}v(\cdot)$ to the resulting a priori (weighted) estimation errors $\sqrt{\bar{\mu}(\cdot)}e_a(\cdot)$ can be expressed in terms of the feedback structure shown in Figure 3.

Define $\gamma(N) = \max_{0\leq i\leq N} \ \mu(i)/\bar{\mu}(i)$ and

$$\Delta(N) \ \triangleq \ \max_{0\leq i\leq N} \left| 1 - f'[\eta(i)]\frac{\mu(i)}{\bar{\mu}(i)} \right|$$

It can also be verified [5] that if $\mu(i)$ is chosen such that $0 < \mu(i)f'[\eta(i)] < 2/\|\mathbf{u}_i\|^2$ then the section shown in Fig. 3 is contractive and leads to an l_2-stable (and, hence, robust) algorithm. Moreover, if $\mu(i)$ is chosen in the middle of the interval specified above, say $\mu_{opt}(i)f'[\eta(i)] = \bar{\mu}(i)$, then the feedback loop is disconnected and the convergence speed is faster. In this case, there will be no energy flowing back into the lower input of the lossless section.

But $\eta(i)$ is still unknown and therefore three suitable approximations for $f'[\eta(i)]$ have been suggested in [5], leading to:

- Choice A: $\mu_{opt}(i) =$

$$\bar{\mu}(i)\min\left(-\frac{1/\beta\ln[1/d(i)-1] + \mathbf{u}_i\mathbf{w}_{i-1}}{d(i) - f[\mathbf{u}_i\mathbf{w}_{i-1}]}, T\right),$$

where T is used as a threshold value in order to prevent large step-sizes.

- Choice B: For $\left(d(i) - \frac{1}{2}\right)\left(f(\mathbf{u}_i\mathbf{w}_{i-1}) - \frac{1}{2}\right) > 0$ we set

$$\mu_{opt}(i) = \frac{2\bar{\mu}(i)}{f'[d(i)] + f'[\mathbf{u}_i\mathbf{w}_{i-1}] + \epsilon}$$

otherwise $\mu_{opt}(i) = \bar{\mu}(i)/f'_{\max}$.

- Choice C:

$$\mu_{opt}(i) = \frac{\bar{\mu}(i)}{\beta\left[\hat{f}[\eta(i)](1-\hat{f}[\eta(i)])\right]+\epsilon}\,, \qquad (4)$$

where ϵ is a small positive constant.

Figure 4 shows the resulting learning curves for a particular simulation, where it is clear that the optimal step-size choices (the two left-most curves) lead to excellent convergence. Extensions to recurrent networks are studied in [6].

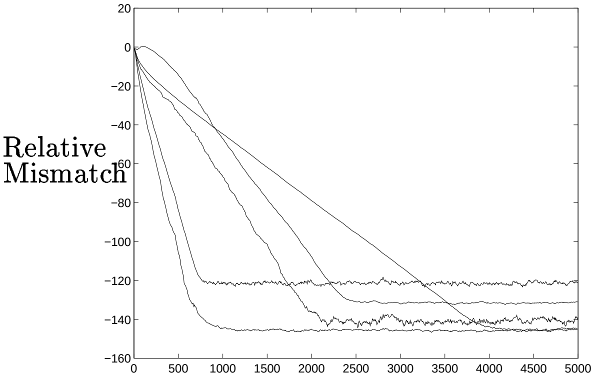


Fig. 4. *Improved convergence for Perceptron training.*

IV. CHANNEL EQUALIZATION

In channel equalization, we are interested in the limiting behaviour of the adaptive scheme (2) as time progresses to infinity. In particular, our objective is to exhibit conditions on $(f, \mu(i))$ under which $e_a(i) \to 0$ as $i \to \infty$ and, consequently, $\hat{z}(i) \to z(i)$ and $\hat{y}(i) \to y(i)$.

We shall assume, without loss of generality, that the update equation (2) is only employed when $e_o(i) \neq 0$ (i.e., we ignore the non-active steps and focus only on updates that involve nonzero error terms $e_o(i)$). In this case, our objective becomes the following: given a sequence of updates with nonzero errors $e_o(i)$, do the resulting weight vector estimates $\mathbf{w}_i$ tend to a value that guarantees $e_a(i) \to 0$ (and, consequently, $e_o(i) \to 0$)?

A. Non-Blind Mode of Operation

The analysis can be carried out by showing that a similar feedback structure to the Perceptron case arises in the context of channel equalization. This is shown in Fig. 5, where h is the function that relates e_a and e_o, $e_o(i) = h[z(i), \hat{z}(i)]e_a(i)$, and is given by

$$h[z(i), \hat{z}(i)] \;=\; \frac{f[z(i)] - f[\hat{z}(i)]}{z(i) - \hat{z}(i)}\,. \qquad (5)$$

If we define

$$\Delta(N) = \max_{0 \le i \le N}\left|1 - \frac{\mu(i)}{\bar{\mu}(i)} h[z(i), \hat{z}(i)]\right|, \;\; \gamma(N) = \max_{0 \le i \le N} \frac{\mu(i)}{\bar{\mu}(i)}\,.$$

Then it can be checked (along the lines of [1], [2]) that if $\Delta(N) < 1$ then the following bounds on the weighted energies of the a priori estimation errors hold:

$$\sqrt{\sum_{i=0}^{N} \bar{\mu}(i)\,|e_a(i)|^2} \le \frac{1}{1-\Delta(N)}\;\|\tilde{\mathbf{w}}_{-1}\|\,. \qquad (6)$$

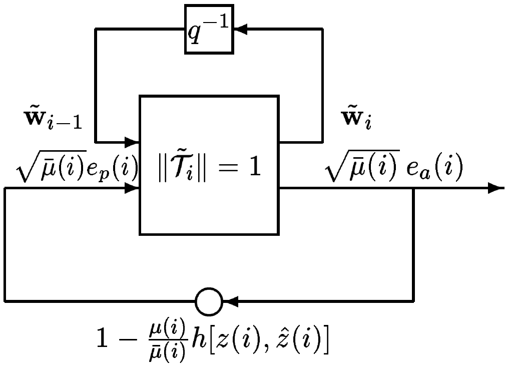

q^{-1}
$\tilde{\mathbf{w}}_{i-1}$
$\tilde{\mathbf{w}}_{i}$
$\sqrt{\bar{\mu}(i)}e_p(i)$
$\|\tilde{\mathcal{T}}_i\| = 1$
$\sqrt{\bar{\mu}(i)}\;e_a(i)$
$1 - \frac{\mu(i)}{\bar{\mu}(i)}h[z(i), \hat{z}(i)]$

Fig. 5. *Structure for non-blind equalization.*

$$\sqrt{\sum_{i=0}^{N} \mu(i)\,|e_a(i)|^2} \le \frac{\gamma^{1/2}(N)}{1-\Delta(N)}\;\|\tilde{\mathbf{w}}_{-1}\|\,. \qquad (7)$$

Relations (6) and (7) are desirable because they imply, when they hold, that in the limit (as $N \to \infty$) the weighted energy of the a priori estimation errors remains bounded or, equivalently, that $\{\sqrt{\mu(i)}\;e_a(i)\}$ and $\{\sqrt{\bar{\mu}(i)}\;e_a(i)\}$ are Cauchy sequences that tend to zero.

The condition $\Delta(N) < 1$ requires (in terms of the real and imaginary parts of h) that

$$\left[1 - \frac{\mu(i)}{\bar{\mu}(i)} h_R(i)\right]^2 \;+\; \frac{\mu^2(i)}{\bar{\mu}^2(i)}\;h_I^2(i) \;<\; 1\,, \qquad (8)$$

which shows that h should necessarily be positive-real. These conditions can be verified for many of the algorithms listed in Table I.

For example, for 2-PSK it can be verified that if $|\mathbf{u}_i\mathbf{w}|$ and $|\mathbf{u}_i\mathbf{w}_{i-1}|$ are uniformly bounded from above, and if $\mu(i)$ is chosen such that $0 < \mu(i) < |\mathbf{u}_i\mathbf{w}_{i-1}|/\|\mathbf{u}_i\|^2$, then $\sqrt{\mu(i)}\;e_a(i) \to 0$ as $i \to \infty$.

Likewise, for the CM algorithm, if $|\mathbf{u}_i\mathbf{w}|$ and $|\mathbf{u}_i\mathbf{w}_{i-1}|$ are uniformly bounded from below and if $\mu(i)$ is chosen such that

$$\left|\frac{\mu(i)}{\bar{\mu}(i)}\;h_I[z(i), \hat{z}(i)]\right| \;\;\text{and}\;\; \left|1 - \frac{\mu(i)}{\bar{\mu}(i)}\;h_R[z(i), \hat{z}(i)]\right|\;,$$

and both less than $1/\sqrt{2}$, where $h[z(i), \hat{z}(i)]$ is evaluated as

$$h[z(i), \hat{z}(i)] = \frac{s(i-D) - \mathbf{u}_i\mathbf{w}_{i-1}|\mathbf{u}_i\mathbf{w}_{i-1}|^2}{s(i-D)|s(i-D)|^{-\frac{2}{3}} - \mathbf{u}_i\mathbf{w}_{i-1}}\,,$$

then we also obtain $\sqrt{\mu(i)}\;e_a(i) \to 0$.

B. Blind Mode of Operation

In the blind mode of operation, the feedback path is modified as shown in Fig. 6, with $(1-h)$ replacing h and where $v_z(i) = f(\mathbf{u}_i\mathbf{w}) - \mathbf{u}_i\mathbf{w}$ denotes the distortion introduced by the channel and by the optimal receiver $\mathbf{w}$.

A contractive map will now require

$$\left|1 - \frac{\mu(i)}{\bar{\mu}(i)}(1 - h[z(i), \hat{z}(i)])\right| < 1 \qquad (9)$$

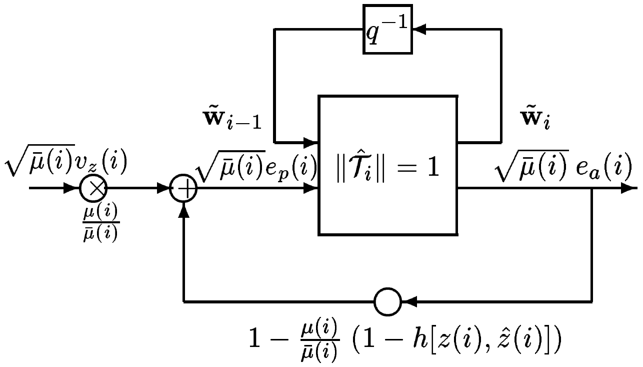

q^{-1}
$\tilde{\mathbf{w}}_{i-1}$
$\tilde{\mathbf{w}}_i$
$\sqrt{\bar{\mu}(i)}v_z(i)$
$\frac{\mu(i)}{\bar{\mu}(i)}$
$\sqrt{\bar{\mu}(i)}e_p(i)$
$\|\hat{\mathcal{T}}_i\| = 1$
$\sqrt{\bar{\mu}(i)}\ e_a(i)$
$1 - \frac{\mu(i)}{\bar{\mu}(i)}\ (1 - h[z(i), \hat{z}(i)])$

Fig. 6. *Structure for blind operation.*

for all possible combinations of $z(i)$ and $\hat{z}(i)$ over the desired interval of time. A necessary condition for this to hold is to require the function $1 - h$ to be positive real. This is in contrast to the non-blind training mode, which requires h itself to be positive real.

One can verify the following for 2-PSK operation. Assume the optimal receiver guarantees $|\mathbf{u}_i\mathbf{w}| = 1$ and its distortion $v_z(\cdot)$ is negligible or has finite energy. If the adaptive weights are only updated whenever $|\mathbf{u}_i\mathbf{w}_{i-1}| > 1$ and if $\mu(i)$ is chosen according to

$$\mu(i) < 2\bar{\mu}(i)\ \frac{|\hat{z}(i)| + 1}{|\hat{z}(i)| - 1}$$

then $\sqrt{\mu(i)}\ e_a(i) \to 0$.

Likewise, assume the weight updates in the normalized CM algorithm are performed only whenever $|\mathbf{u}_i\mathbf{w}_{i-1}| > 1 + \epsilon$ for some given positive $\epsilon \ll 1$. Assume also that the optimal receiver guarantees $|\mathbf{u}_i\mathbf{w}| < 1 + \epsilon$. If $\mu(i) < 2\bar{\mu}(i)$, and if the optimal channel-receiver distortion is negligible or has finite energy, then we also obtain $\sqrt{\mu(i)}\ e_a(i) \to 0$.

V. SIMULATION RESULTS FOR CHANNEL EQUALIZATION

The channel employed in all simulations was $C(q^{-1}) = 1 + 0.9q^{-1}$, and the receiver length was taken as $M = 3$.

We consider first the non-blind mode of operation. The step-size parameter was chosen in two ways: a non-normalized mode where $\mu(i) < \bar{\mu}(i)$ (as is the case with standard gradient algorithms [2]) and a normalized mode where $\mu(i) < \bar{\mu}(i)|\hat{z}(i)|$ as suggested by the discussion at the end of Sec. 4.1. Fig. 7 depicts the results for the two modes, where α denotes $\mu(i)/\bar{\mu}(i)$. The figure shows the Bit-Error-Rates (BER); the ratio of falsely detected bits to the overall transmitted bits. The algorithms were run for $N = 200$ steps and the results averaged over 20 Monte Carlo runs.

We now consider the blind mode of operation. As suggested by the discussion in Sec. 4.2, a convergent (and robust) performance in the blind mode of operation can be guaranteed as long as the operation of the adaptive equalizer is restricted to "large" enough values of $\hat{z}(i)$. To verify this fact, we ran several simulations with $\alpha = 0.1 = \mu(i)/\bar{\mu}(i)$ and for different values of β (where β is used to determine when to update the weight estimates, viz., for $|\hat{z}(i)| > \beta$). Fig. 8 depicts BER values as a function of β. For $\beta = 0$, the standard 2-PSK algorithm is obtained. The experiment was run for $N = 200$ steps and the results averaged over 20 runs. For values of $\beta > 0$, the algorithm shows a considerably improved behaviour. However, the larger the β, the smaller becomes the improvement, since then the updates become less frequent.

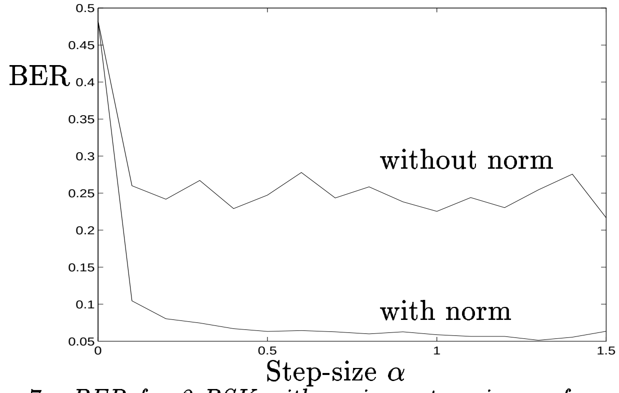


Fig. 7. *BER for 2-PSK with various step-sizes α for normalized and non-normalized mode.*

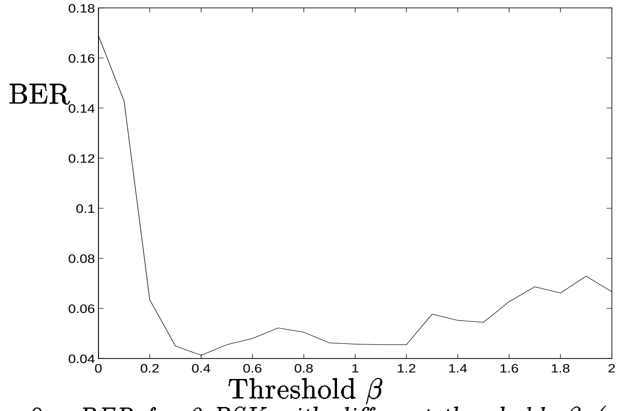


Fig. 8. *BER for 2-PSK with different thresholds β ($\alpha = 0.1$).*

The next experiment is for the normalized CM algorithm. Fig. 9 depicts three averaged learning curves obtained from averaging $|e(i)|^2 = |s(i-D) - \hat{z}(i)/|\hat{z}(i)|^2$ over 200 runs. It shows that the normalized CMA has superior performance if it is not updated at every time instant.

References

[1] A. H. Sayed and M. Rupp, "Error energy bounds for adaptive gradient algorithms," *IEEE Transactions on Signal Processing*, Aug. 1996.

[2] M. Rupp and A. H. Sayed, "A time-domain feedback analysis of filtered-error adaptive gradient algorithms," *IEEE*

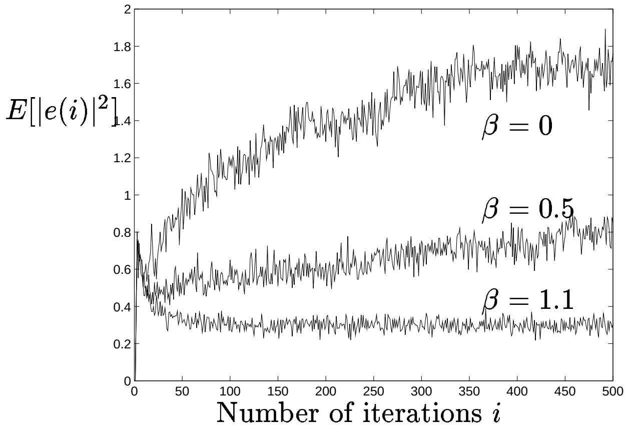


Fig. 9. *Instantaneous error energy for normalized CM algorithm with step-size $\alpha = 1$ and threshold $\beta = 0, 0.5, 1.1$.*

Trans. on Signal Processing, July 1996. See also *Proc. SPIE*, vol. 2563, pp. 458-469, San Diego, July 1995.

[3] S. Haykin, *Neural Networks: A Comprehensive Foundation*, MacMillan Publishing Company, 1994.

[4] S. Haykin, *Blind Deconvolution*, NJ: Prentice Hall, 1994.

[5] A.H. Sayed, M. Rupp, “A feedback analysis of Perceptron learning for neural networks,” *Proc. of Asilomar Conference,* October 1995.

[6] M. Rupp and A.H. Sayed, “On the robustness of Perceptron learning recurrent networks,” *Proc. 13th IFAC World Congress*, San Francisco, June 1996.