LOG ADAPTIVE FILTERS - STRUCTURES AND ANALYSIS FOR THE SCALAR CASE

M. Rakijas and N. J. Bershad

Electrical and Computer Engineering, School of Engineering
University of California, Irvine
Irvine, CA 92697, USA

ABSTRACT

Feed-forward multi-layer neural networks (MLNN's) are complex nonlinear learning systems which can be trained by well-known rules such as back-propagation (BP). The resulting adaptation procedures are extremely difficult to analyze for stochastic training data. Significant analytic results have been obtained for the single-layer case and for some simple two-layer cases. Recently, a structural simplification has been studied which models each threshold function as a linear device. This linearized MLNN can only create hyperplane decision rules after convergence. However, the multiplicative behavior of the layers may offer some performance advantages over linear adaptive algorithms (LMS or RLS) when used for a linear problem. A new log-domain linear MLNN adaptive structure is proposed and analyzed here. The log operation converts the layer multiplications into additions whereupon linear analysis techniques can be used. The transient and steady-state statistical behavior of the log linear MLNN is analyzed for Gaussian training data. Deterministic recursions are derived for the mean and fluctuation behavior of the new algorithm. These recursion are shown to be in excellent agreement with Monte Carlo simulations.

1. INTRODUCTION

The MLNN is denoted a linear MLNN for linear output neurons. However, the linear MLNN is not a linear device because of the layer multiplication. The BP learning behavior of the linear MLNN has been analyzed in [1] for some special symmetrical cases. Convergence speed and mean square error (MSE) was studied. The analysis in [1] suggested that the linear MLNN performs better (i.e. a smaller misadjustment error for the same transient response) than the LMS algorithm under some operating conditions. The convergence speed depends on the number of layers in the linear MLNN. First order analysis of the log of the weights yielded reasonable theoretical predictions for the convergence rate and MSE. The first order approximations were extended in [2].

1.1 Adaptation in the Log Domain

A structure was selected which converts the multiplicative operations of the linear MLNN layers to additive operations. A single tap version of the new structure is shown in Figure 1. The desired output is generated by multiplying the input by a real scalar *a*. The new adaptive structure estimates the unknown system by

$$\hat{d}[n] = x[n]\exp\left(\mathbf{1}^T w[n]\right) \tag{1}$$

where $w^T = [w_1 \quad w_2 \quad \cdots \quad w_L]$; $\mathbf{1}^T = [1 \quad 1 \quad \cdots \quad 1]$.

The input process x[n] is a real zero-mean, white Gaussian sequence (ZMWGN) with unity variance. The independent additive observation noise r[n] is also a real ZMWGN process with unity variance. The gradient descent weight adaptation is

$$w[n+1] = w[n] + \mu x[n]\exp\left(w^{\dagger}\mathbf{1}\right)e[n]\mathbf{1} \tag{2}$$

where † signifies complex conjugate transpose.

1.2 Multi-Layered - Real Initialization

Weight initialization is a primary issue because of the logarithm operation. The coefficients are real if they are initialized real and d, x are real. This effect simplifies the analysis. Thus, the real initialization case is examined first. The gradient of $J = E\left(e^* e\right)$ is

$$\nabla J = -2R_{xd}\exp\left(\mathbf{1}^{\mathrm{T}} w\right) + 2R_x \tag{3}$$

When the weight vector is fixed at the optimum value, say w_{opt}, the gradient of the MSE is zero and

$$\exp\left(\mathbf{1}^T w_{opt}\right) = R_x^{-1}R_{xd} \tag{4}$$

The right side of (4) must be greater than zero for real adaptive weights to converge to w_{opt}. Equation (4) also shows that an infinite number of weight sets achieve the optimum. For the unknown system, $d = ax + r$ and $R_x^{-1}R_{xd} = a$. Therefore, a>0. If $R_x^{-1}R_{xd}<0$ and the weights are initialized real, (1) cannot converge to w_{opt} in any sense.

Mean Recursion

The mean behavior of (2) is given by the recursion

$$E\left(w[n+1]\right) = E\left(w[n]\right) + \mu E\left[\exp\left(\mathbf{1}^T w[n]\right)xe\right]\mathbf{1} \tag{5}$$

The stationary point occurs when E(w[n+1])=E(w[n]) or when

$$R_{xd}E\left[\exp\left(\mathbf{1}^T w[n]\right)\right] = R_x E\left[\exp\left(2\left(\mathbf{1}^T w[n]\right)\right)\right] \tag{6}$$

To proceed further with the analysis, the weights are assumed Gaussian. Then, the expectations in (6) can be evaluated using the joint characteristic function for multi-variate Gaussian random variables. For a real Gaussian random vector, this is

$$\Phi_w(w) = E\left(j\Omega^T w\right) = \exp\left(j\Omega^T \overline{w}\right)\exp\left(-\frac{1}{2}\Omega^T C\Omega\right) \tag{7}$$

Hence

$$E\left[\exp\left(\alpha\left(\mathbf{1}^T w\right)\right)\right] = \exp\left(\alpha\left(\mathbf{1}^{\mathbf{T}}\overline{w}\right)\right)\exp\left(\frac{\alpha^2}{2}\mathbf{1}^T C_w\mathbf{1}\right) \tag{8}$$

where α is a real scalar,

$$C_w = E\left[(w-\overline{w})(w-\overline{w})^T\right] \quad \text{and} \quad \mathrm{E}(\mathrm{w}) = \overline{w} \tag{9}$$

Using (8) in (6) yields

$$R_x^{-1}R_{xd} = \exp\left(\mathbf{1}^T\overline{w}\right)\exp\left(\frac{3}{2}\mathbf{1}^T C_w\mathbf{1}\right) \tag{10}$$

The stationary point can be reached since $R_x^{-1}R_{xd}$ is positive.

Variance recursion

A recursion for C_W is needed to evaluate (10). Using (2) and (5),

$$\begin{aligned}\tilde{w}[n+1] &\equiv w[n+1]-\overline{w}[n+1] \\ &= \tilde{w}[n]+\mu\left\{\left[xd\exp\left(\mathbf{1}^T w[n]\right)-R_{xd}E\left(\exp\left(\mathbf{1}^T w[n]\right)\right)\right]\right. \\ &\quad \left.+\left[x^2\exp\left(2\left(\mathbf{1}^T w[n]\right)\right)-R_x E\left(\exp\left(2\left(\mathbf{1}^T w[n]\right)\right)\right)\right]\right\}\mathbf{1}\end{aligned} \tag{11}$$

Note that $\tilde{w}$ is a zero mean vector. Using (11) in (9) yields

$$C_w[n+1] = C_w[n]+\mu\left(A_x+A_{xd}\right)+\mu^2 B \tag{12}$$

where

$$A_x = R_x\exp\left(2\left(\mathbf{1}^T\overline{w}\right)\right)\left\{E\left[\exp\left(2\left(\mathbf{1}^T\tilde{w}\right)\right)\left(\tilde{w}\mathbf{1}^T+\mathbf{1}\tilde{w}^T\right)\right]\right\} \tag{13}$$

$$A_{xd} = R_{xd}\exp\left(\mathbf{1}^T\overline{w}\right)\left\{E\left[\exp\left(\mathbf{1}^T\tilde{w}\right)\left(\tilde{w}\mathbf{1}^T+\mathbf{1}\tilde{w}^T\right)\right]\right\} \tag{14}$$

and B is the coefficient of the μ^2 term. With α a real scalar, and

$$E\left[\tilde{w}\exp\left(\alpha\left(\mathbf{1}^T\tilde{w}\right)\right)\right] = \alpha\left(C_w\mathbf{1}\right)\exp\left(\frac{\alpha^2}{2}\mathbf{1}^T C_w\mathbf{1}\right) \tag{15}$$

the form of the expectations in A_x, A_{xd} and B can be written as

$A_x = -2R_x\exp\left(2\left(\mathbf{1}^T\overline{w}+\mathbf{1}^T C_w\mathbf{1}\right)\right)\left(C_w\mathbf{1}\mathbf{1}^T+\mathbf{1}\mathbf{1}^T C_w\right)$,

$A_{xd} = R_{xd}\exp\left(\mathbf{1}^T\overline{w}+\frac{1}{2}\mathbf{1}^T C_w\mathbf{1}\right)\left(C_w\mathbf{1}\mathbf{1}^T+\mathbf{1}\mathbf{1}^T C_w\right)$, and

$$\begin{aligned}B = &\left\{R_x R_{xd}\exp\left(2\left(\mathbf{1}^T\overline{w}+\mathbf{1}^T C_w\mathbf{1}\right)\right)+R_{xd}^2\exp\left(2\left(\mathbf{1}^T\overline{w}\right)+\mathbf{1}^T C_w\mathbf{1}\right)\left(2\exp\left(\mathbf{1}^T C_w\mathbf{1}\right)-1\right)\right. \\ &-R_x R_{xd}\exp\left(3\left(\mathbf{1}^T\overline{w}\right)+\frac{5}{2}\mathbf{1}^T C_w\mathbf{1}\right)+\left(6\exp\left(2\left(\mathbf{1}^T C_w\mathbf{1}\right)\right)-2\right) \\ &\left.-R_x^2\exp\left(4\left(\mathbf{1}^T\overline{w}\right)+4\left(\mathbf{1}^T C_w\mathbf{1}\right)\right)+\left(3\exp\left(4\left(\mathbf{1}^T C_w\mathbf{1}\right)\right)-1\right)\right\}\mathbf{1}\mathbf{1}^T\end{aligned} \tag{16}$$

The form of B in (16) is significant. B is a scalar multiplied by a ones dyadic matrix. C[1] is the zero matrix since the weights are initialized to a constant. Thus, to satisfy (5) and (12), the covariance matrix elements must be identical.

$$\left(C_w[n]\right)_{ij} = \left(C_w[n]\right)_{kl} \qquad \forall i,j,k,l = 1,2,\ldots,L;\ n \geq 1 \tag{17}$$

The vector $C_W\mathbf{1}$ is an L-dimensional vector with Lc as its unique element, and $1^T C_W\mathbf{1}=L^2c$. Thus, the covariance matrix recursion reduces to a scalar recursion. Let

$$a_x = \left(A_x\right)_{ij};\ a_{xd} = \left(A_{xd}\right)_{ij};\ b = \left(B\right)_{ij} \tag{18}$$

An expression can be found for the stationary point of (12). The stationary point is defined by $\mu\left(a_x+a_{xd}\right) = -\mu^2 b$ (19)

or

$$\mu = \frac{2Lc\exp\left(\mathbf{1}^T\overline{w}+\frac{L^2c}{2}\right)\left[2R_x\exp\left(\mathbf{1}^T\overline{w}+\frac{3}{2}L^2c\right)-R_{xd}\right]}{b} \tag{20}$$

Equation (10) relates the mean weight, the covariance matrix and $R_x^{-1}R_{xd}$. Using (10) in (20) yields an expression for the weight covariance as a function of μ:

$$\mu = \frac{2Lc}{R_x^{-1}R_{xd}\left[3\exp\left(3L^2c\right)-6\exp\left(L^2c\right)+2\right]+R_d} \tag{21}$$

Equation (20) determines MSE as a function of μ. As a function of c, the MSE is

$$\begin{aligned}J = E\left(e^2\right) &= R_d-2R_{xd}E\left(\exp\left(\mathbf{1}^{\mathbf{T}}w\right)\right)+R_x E\left(\exp\left(2\left(\mathbf{1}^T w\right)\right)\right) \\ &= R_d-R_x^{-1}R_{xd}^2\exp\left(-L^2c\right)\end{aligned} \tag{22}$$

The minimum MSE occurs for constant w_{opt}. Using (4) in (22)

$$J_{opt} = R_d-R_x^{-1}R_{xd}^2 \tag{23}$$

Thus, the excess MSE, J_{exc}, is

$$J_{exc} = R_x^{-1}R_{xd}^2\left(1-\exp\left(-L^2c\right)\right) \tag{24}$$

1.3 Reduction to Single Layer Equivalents

Suppose a realization yields sequences x[n], r[n] and d[n], a fixed *a* with weight sum initialization, ($\mathbf{1}^{\mathrm{T}}$w[1]). Consider two adaptive systems with different weight sets but the same output $\hat{d}[n]$ for all n>0. We denote the two systems indistinguishable. The MSE in (22) is constant for constant L^2c. L^2c in (21) is constant if μL constant. Consider two systems in a family defined by constant μL. They have the same structure, except for the number of layers. Suppose system 1 within the family has L_1 layers, a specific μ_1 and has achieved weight sum at time n of $s_n=\mathbf{1}^{\mathrm{T}}w[n]$. Consider system 2 with L_2 layers with weight sum s_n. Both systems yield the same $\hat{d}[n]$ because of the same weight sum. For each system, the weight sum update of (1) is

$$\mathbf{1}^T\left(w[n+1]-w[n]\right) = \mu_i L_i x[n]\exp\left(\mathbf{1}^T w[n]\right)e[n] \tag{25}$$

which is independent of i. Both systems have the same weight sum for time n+1 and thus have the same $\hat{d}[n+1]$. Inductively, systems 1 and 2 are indistinguishable. Thus systems within the family defined by constant μL are indistinguishable. Therefore, it is sufficient to study the behavior of a one layer system since it defines performance of the multi-layered systems of the family.

1.4 Complex Initialization

In general, the sign of *a* is not known a priori. Thus, the behavior of (2) must also be studied when the weights are complex. Since the learning rule remains scalar, complex weight values do not change the single versus multiple layer equivalence. Thus, the multi-layered system behavior can be described by the single layer system for the family behavior.

Real and Imaginary Parts

Equation (2) can be re-written in terms of separate recursions for the real and imaginary parts:

$$\begin{aligned}w_R[n+1] &= w_R[n]+\mu\left[xd\exp\left(w_R[n]\right)\cos\left(w_I[n]\right)-x^2\exp\left(2w_R[n]\right)\right] \\ w_I[n+1] &= w_I[n]+\mu\left[xd\exp\left(w_R[n]\right)\sin\left(w_I[n]\right)\right]\end{aligned} \tag{26}$$

where w_R and w_I are the real and imaginary parts of w, respectively. Note that the recursions in (26) are not symmetric.

Mean Recursions

The analysis of the complex mean recursions proceeds similarly to that for the real mean recursions. Averaging (26), yields

$$\begin{aligned}\overline{w}_R[n+1] &= \overline{w}_R[n] + \mu\Big[R_{xd}E\big[\exp(w_R[n])\cos(w_I[n])\big] \\ &\quad - R_x\exp\big(2(\overline{w}_R[n]+C_R[n])\big)\Big] \\ \overline{w}_I[n+1] &= \overline{w}_I[n] + \mu\Big[R_{xd}E\big[\exp(w_R[n])\sin(w_I[n])\big]\Big]\end{aligned} \tag{27}$$

Equation (27) requires both the variances and cross-covariance of the real and imaginary parts of the weight vector. Thus, recursions must be derived for

$$\begin{aligned}C_R[n+1] &\equiv E\big[\tilde{w}_R[n+1]\tilde{w}_R[n+1]\big] \\ C_I[n+1] &\equiv E\big[\tilde{w}_I[n+1]\tilde{w}_I[n+1]\big] \\ C_{RI}[n+1] &\equiv E\big[\tilde{w}_R[n+1]\tilde{w}_I[n+1]\big]\end{aligned} \tag{28}$$

These terms can be evaluated using the characteristic function of complex non-circular symmetric Gaussian random variables,

$$E\big[\exp(\alpha w_R + j\beta w_I)\big] \quad \text{and} \quad E\big[\exp(\alpha w_R - j\beta w_I)\big]$$

for real jointly Gaussian random variables w_R and w_I with arbitrary covariance. We evaluate using (7), so that

$$\begin{aligned}E\big[\exp(\alpha w_R + j\beta w_I)\big] &= \exp(\alpha\overline{w}_R + j\beta\overline{w}_I)\exp\left(\frac{1}{2}\left(\alpha^2 C_R + 2j\alpha\beta - \beta^2 C_I\right)\right) \\ E\big[\exp(\alpha w_R - j\beta w_I)\big] &= \exp(\alpha\overline{w}_R - j\beta\overline{w}_I)\exp\left(\frac{1}{2}\left(\alpha^2 C_R - 2j\alpha\beta - \beta^2 C_I\right)\right)\end{aligned} \tag{29}$$

After some algebra, the mean recursions are

$$\begin{aligned}\overline{w}_R[n+1] &= \overline{w}_R[n] + \mu\exp\left(\overline{w}_R + \frac{1}{2}C_R\right) \\ &\quad \bullet\left[R_{xd}\exp\left(-\frac{1}{2}C_I\right)\cos(\overline{w}_I + C_{RI}) - R_x\exp\left(\overline{w}_R + \frac{3}{2}C_R\right)\right] \\ \overline{w}_I[n+1] &= \overline{w}_I[n] + \mu R_{xd}\exp\left(\overline{w}_R + \frac{1}{2}(C_R - C_I)\right)\sin(\overline{w}_I + C_{RI})\end{aligned} \tag{30}$$

Fluctuation Recursions

Using (29) in (28) yields the recursions for C_R, C_I and C_{RI} and are given in [2].

Optimum Weights and the Minimum MSE

The gradient of the MSE surface is zero when $w=w_{opt}$. The gradient is zero when

$$\sin\left(w_{I_{opt}}\right) = 0 \tag{31}$$

and

$$R_x^{-1}R_{xd}\cos\left(w_{I_{opt}}\right) = \exp\left(w_{R_{opt}}\right) \tag{32}$$

Equation (31) holds because the unknown system is assumed real. Using (31) and (32),

$$w_{R_{opt}} = \ln\left|R_x^{-1}R_{xd}\right|, \quad w_{I_{opt}} = \begin{cases}0,\pm 2\pi,\pm 4\pi\ldots & R_x^{-1}R_{xd} > 0 \\ \pm\pi,\pm 3\pi\ldots & R_x^{-1}R_{xd} < 0\end{cases} \tag{33}$$

or equivalently

$$\exp\left(w_{opt}\right) = \exp\left(w_{opt}^*\right) = R_x^{-1}R_{xd} \tag{34}$$

The same minimum MSE occurs as in the real case.

Stationary Points

At least two stationary points are evident for the mean (27) and covariance recursions. One stationary point is similar to the real case and occurs when

$$C_I = C_{RI} = 0 \tag{35}$$

and

$$\cos(\overline{w}_I)\exp\left(\overline{w}_R + \frac{3}{2}C_R\right) = R_x^{-1}R_{xd} \tag{36}$$

Real weights occur when $\sin(w_I[n])=0$ for all n. If this stationary point is reached, the MSE of the complex log adaptive filter can approach the minimum MSE. As in the real case, (21) relates μ and C_R for L=1. The second stationary point is approached if the unknown system is unreachable. The adaptive increments approach zero as $\overline{w}_R \to -\infty$,

$$\left.\begin{aligned}\overline{w}_R[n+1] &\approx \overline{w}_R[n] \\ \overline{w}_I[n+1] &\approx \overline{w}_I[n]\end{aligned}\right| \begin{aligned}C_R[n+1] &\approx C_R[n] \\ C_I[n+1] &\approx C_I[n] \\ C_{RI}[n+1] &\approx C_{RI}[n]\end{aligned}\Bigg\} \text{ as } \overline{w}_R \to -\infty \tag{37}$$

2. MONTE CARLO SIMULATIONS

Figure 2 shows c(μ) of equation (21) with L as a parameter. The curves can be used to choose μ for a desired MSE. Equation (22) converts c to MSE. Several Monte Carlo simulations were made to support the theory. 10 runs were averaged for each simulation result. The first set is similar to the real case with parameters a=7, μ=.003, R_x=1, R_r=1 and R_d=50.

Weight initialization significantly affects the performance (multiple stationary points) and requires further study. Adaptive systems are customarily initialized at rest. For example, LMS weights are initialized at zero and the system output is zero at time zero. With log adaptive filters, if the weights are initialized at zero, then $\exp(\mathbf{1}^T w)=1$ and the adaptive system is not at rest. There is no finite initial value for which the system is at rest and if the weights approach -∞, the update in (2) is made arbitrarily small turning the system off. Similarly, the initial learning update can be made arbitrarily large. As a compromise, the weights are initialized so that $\exp(\mathbf{1}^T w)=1$. To exercise the algorithm's imaginary part, the weight was initialized at $w[1]=.1j$. Figure 3 shows the theoretical and simulated excess MSE. Figure 4 shows C_R, C_I and C_{RI} for theory and simulations. Figure 5 suggests that this small imaginary part initialization is not sufficient for identifying negative systems. The same μ and w[1] was used, only a=-7. In this case, $\exp(w_R)\to 0$ implying that the weights are approaching the second stationary point, i.e. $w_R \to -\infty$. This behavior does not converge. Figure 6 shows the correctly identified negative system with $w[1]=j$. The imaginary part is sufficient to allow convergence to the optimal weight. Not shown is that a=7 can also be identified with this initialization and that the theory and MC simulations agree well for the mean weight.

3. RESULTS AND CONCLUSIONS

The algorithm in (26) identified 1) a positive scalar for the initial complex weight in Fig. 3, 2) a positive or negative scalar for the complex initial weight in Fig. 6. The algorithm in (26) could not identify the negative scalar for the initial complex weight in Fig. 5. The MSE surface is not a quadratic function of the weights and may have multiple minima satisfying $\nabla J = 0$. Hence, the global minimum of the surface may not be accessed from every initial state. Thus, there are initialization regions for which (26) can identify positive scalars, negative or both. This will be examined more closely in a future work.

4. REFERENCES

[1] M Ibn Kahla, Z. Faraj, F. Castanie, J. C. Hoffman, "Multi-layer adaptive filters trained with back propagation: A statistical

approach", Signal Processing, 40(1994), pp. 65-85
[2] M. Rakijas, unpublished work
[3] Papoulis, A. Probability, Random Variables, and Stochastic Processes. New York: McGraw Hill, Inc., 1991

5. FIGURES

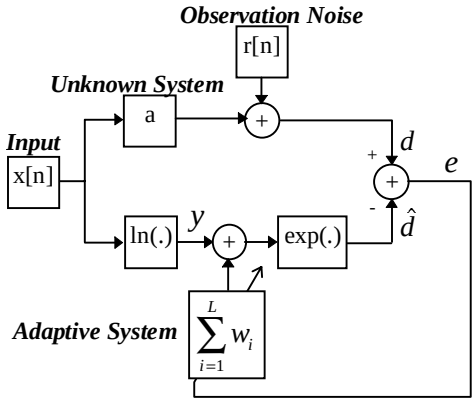

Figure 1 - Block Diagram of Log Adaptive Filter

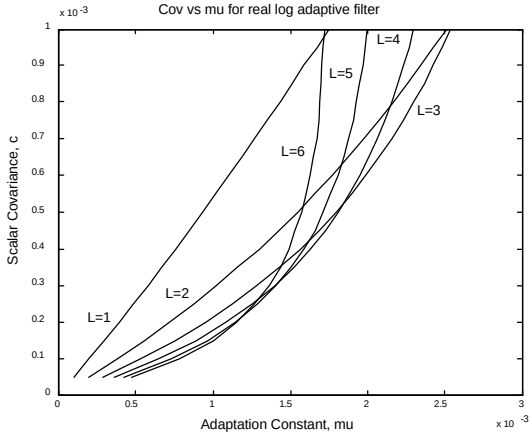

Figure 2 - Weight variance as a function of learning rate

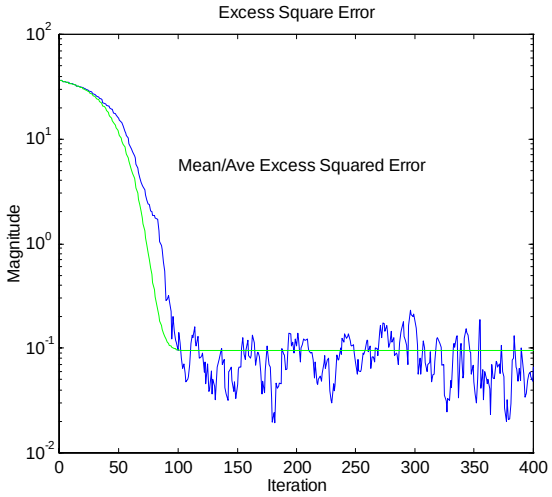

Figure 3 - Excess error of complex algorithm for a=7

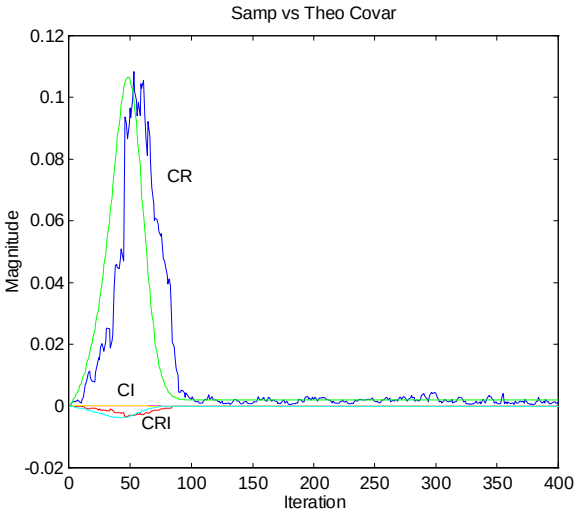

Figure 4 - Variances and Covariances for a=7

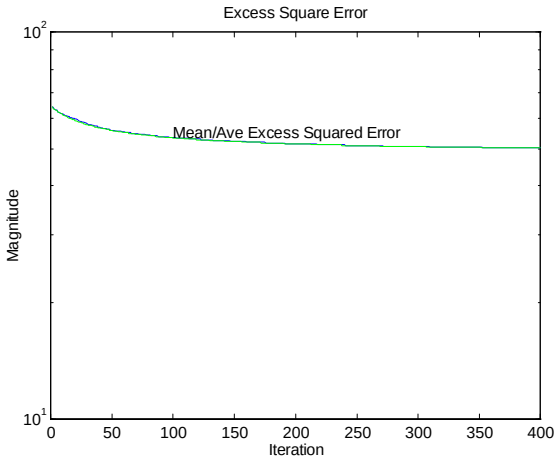

Figure 5 - Mean and average excess error for a=-7

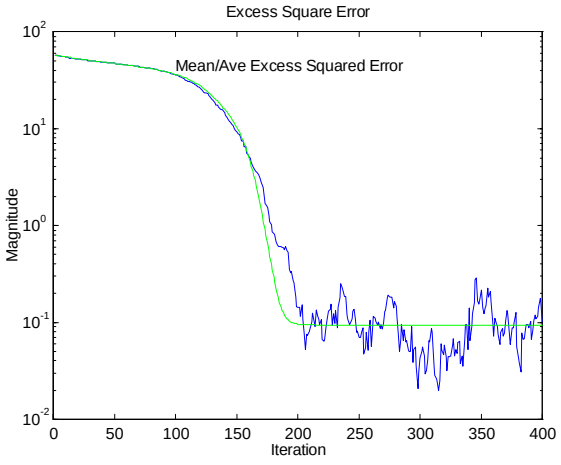

Figure 6 - Excess error for a=-7 with new initialization