OPTIMAL CONVERGENCE FACTOR FOR THE LMS/NEWTON ALGORITHM

Paulo S. R. Diniz and Luiz W. P. Biscainho

Programa de Engenharia Elétrica
COPPE/EE/Federal University of Rio de Janeiro
C. P. 68504 Cep. 21945 Rio de Janeiro, RJ Brazil

ABSTRACT
An efficient approach for the computation of the optimum convergence factor for LMS/Newton algorithm applied to transversal FIR structure is proposed. The approach leads to a variable step size algorithm that results in a dramatic reduction in convergence time. The algorithm is evaluated in systems identification applications.

I. INTRODUCTION
Recently a number of publications have proposed alternative variable step sizes methods to be employed in least mean square (LMS) algorithms in order to improve its convergence rate [1]–[6]. These approaches are in general efficient when the eigenvalue spread of the input signal autocorrelation matrix is moderate.

When the input signal autocorrelation matrix has widely spread eigenvalues, more complex algorithms are required in order to achieve convergence at a reasonable speed. One approach is to use the transform domain adaptive filters [7]–[8], where the input signal vector is first processed through an orthogonal transform and the filter taps are updated through the normalized LMS algorithm. The success of this approach depends on a deep knowledge of the input signal statistics [8].

Another approach is to use a fast convergence algorithm such as the LMS/Newton [9], that requires an appropriate choice of the convergence factor. In this paper, an optimum instantaneous convergence factor for this type of algorithm is first proposed that yields fast convergence and requires no information about the input signal statistics. The approach leads to a variable step size LMS/Newton algorithm applied to FIR transversal filter. Alternative algorithms for IIR filters are possible but they are quite different from the one proposed here [10].

Simulations in system identification applications are included to confirm the usefulness of the proposed approach.

II. SQUARE ERROR VARIATION
Consider the adaptive FIR filter shown in Fig. 1. The output signal is given by

\[ y_k = \sum_{i=0}^{N} w_{ik} x_{k-i} = W_k^T X_k \]  

(1)

where \( W_k \) and \( X_k \) denote the tap weights vector and the input signal vector respectively.

The output error is given by

\[ e_k = d_k - y_k \]  

(2)

The tap weights are adapted in order to minimize the Mean Square Error (MSE) defined as follows

\[ \text{MSE} = \mathbb{E}[e_k^2] \]  

(3)

The square error can be calculated by

\[ e_k^2 = d_k^2 + W_k^T X_k X_k^T W_k - 2 W_k^T d_k x_k \]  

(4)

The MSE that would result if the filter coefficients were fixed at \( W_k \) is given by

\[ \xi_k = \mathbb{E}[e_k^2] = E[d_k^2] + W_k^T R W_k - 2 W_k^T P \]  

(5)

where \( R = E[X_k X_k^T] \) is the input correlation matrix, and \( P = E[d_k X_k] \) is the cross correlation between the desired response and the input signal vector.

From (5) it follows that

\[ \xi_{k+1} = E[d_{k+1}^2] + W_{k+1}^T R W_{k+1} - 2 W_{k+1}^T P \]  

(6)

By considering \( W_{k+1} = W_k + \Delta W_k \), it follows that

\[ \xi_{k+1} = E[d_{k+1}^2] + (W_k + \Delta W_k)^T R (W_k + \Delta W_k) - 2(W_k + \Delta W_k)^T P = E[d_{k+1}^2] + W_k^T R W_k - 2 W_k^T P + W_k^T R \Delta W_k + \Delta W_k^T R W_k + \Delta W_k^T R \Delta W_k - 2 \Delta W_k^T P \]  

(7)

Since \( W_k^T R \Delta W_k = \Delta W_k^T R W_k \) due to the symmetry property of \( R \), and using (4), the equation above can be rewritten as

\[ \xi_{k+1} = \xi_k + \Delta W_k^T R W_k + \Delta W_k^T R \Delta W_k - 2 \Delta W_k^T P \]  

(8)

The gradient vector of (4) is given by

\[ \nabla \xi_k = 2 R W_k - 2P \]  

that by rearranging yields

\[ \xi_{k+1} = \xi_k + \Delta W_k^T R W_k + \Delta W_k^T R \Delta W_k - 2 \Delta W_k^T P \]  

(8)
\[ P = \frac{V}{R} + W_k \]  

From (8), (9) and by defining \( \Delta \xi_k \equiv \hat{\xi}_{k+1} - \xi_k \) it follows that

\[ \Delta \xi_k = \Delta W_k^T \hat{P} + \Delta W_k^T R \Delta W_k. \]  

The objective in order to increase the convergence rate is to reduce \( \Delta \xi_k \) as much as possible, since in general \( \Delta \xi_k < 0 \).

III. VARIABLE STEP SIZE METHOD FOR THE LMS/NEWTON ALGORITHM

The tap weights updating in the LMS/Newton algorithm is performed as follows

\[ W_{k+1} = W_k - \mu \left( \hat{R}^{-1} \right)_k \hat{v}_k \]  

where

\[ \hat{v}_k = -2 \xi_k X_k \]  

represents an instantaneous estimate of the gradient vector and \( \left( \hat{R}^{-1} \right)_k \) is the inverse of an estimate of the input correlation matrix.

The estimate \( \left( \hat{R}^{-1} \right)_k \) is obtained through the following recursion [9],

\[ \left( \hat{R}^{-1} \right)_k = \frac{1}{1 - \alpha} \left( \left( \hat{R}^{-1} \right)_{k-1} - \frac{\left( \left( \hat{R}^{-1} \right)_{k-1} X_k \right)^T}{1 - \alpha} \left( \left( \hat{R}^{-1} \right)_{k-1} X_k \right) \right) \]  

that is obtained by applying the matrix inversion lemma [9] to an estimate of \( R \) given by

\[ \hat{R}_k = (1 - \alpha) \hat{R}_{k-1} + \alpha X_k X_k^T \]  

where \( 0 < \alpha < 1 \).

It should be noted that the convergence factor \( \mu \) in (11) must be in the range

\[ 0 < \mu < \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} \]  

in order to guarantee the algorithm convergence [9], where \( \lambda_{\text{max}} \) and \( \lambda_{\text{min}} \) are respectively the maximum and average values of the eigenvalues of \( R \).

Suppose now that the convergence factor \( \mu \) is made variable with the objective of achieving a faster convergence rate, thus

\[ \mu_k = b \alpha_k \]  

where \( b \) is a fixed factor and \( \alpha_k \) is variable factor. In this case,

\[ \Delta W_k = W_{k+1} - W_k = 2b \alpha_k \left( \hat{R}^{-1} \right)_k \xi_k X_k \]  

By replacing (11) and (17) in (10), and considering that \( \left( \hat{R}^{-1} \right)_k \) is symmetric, it follows that

\[ \Delta \xi_k = \left[ 2b \alpha_k \left( \hat{R}^{-1} \right)_k X_k \right]^T \hat{P} \]  

\[ + \left[ 2b \alpha_k \left( \hat{R}^{-1} \right)_k X_k \right]^T R \hat{R}_k \left[ 2b \alpha_k \left( \hat{R}^{-1} \right)_k X_k \right] = \]  

\[ \left[ -4b \alpha_k \xi_k^T \left( \hat{R}^{-1} \right)_k X_k \right]^T \]  

\[ + 4b \alpha_k \xi_k^T \left( \hat{R}^{-1} \right)_k X_k \hat{R}_k \left( \hat{R}^{-1} \right)_k X_k = \]  

\[ \left[ -4b \alpha_k \xi_k^T \left( \hat{R}^{-1} \right)_k X_k \right]^T \]  

\[ + 4b \alpha_k \xi_k^T \left( \hat{R}^{-1} \right)_k X_k \hat{R}_k \left( \hat{R}^{-1} \right)_k X_k \]  

From (9), it follows by taking the derivative related to \( W \) that

\[ R = \frac{1}{2} \frac{\partial P}{\partial W} \]  

Since the gradient estimate was calculated by (12), it is reasonable to consider that the most consistent instantaneous estimate of \( R \) is given by

\[ \hat{R}_k = \frac{1}{2} \frac{\partial \hat{P}}{\partial W} = \frac{1}{2} \frac{\partial \hat{P}}{\partial W} \]  

\[ = X_k \hat{X}_k^T \]  

In this case, from (18) it follows that

\[ \Delta \xi_k = -4b \alpha_k \xi_k^T \left( \hat{R}^{-1} \right)_k X_k + 4b \alpha_k \xi_k^T \left( \hat{R}^{-1} \right)_k X_k \hat{R}_k \left( \hat{R}^{-1} \right)_k X_k \]  

If the same variable step size is used in (13) to estimate \( \hat{R}^{-1} \), it can be shown that

\[ \left( \hat{R}^{-1} \right)_k X_k = \hat{X}_k \left[ \frac{1}{1 - \alpha_k} - \frac{\left( \left( \hat{R}^{-1} \right)_k X_k \right)^T}{1 - \alpha_k} \right] \left( \hat{R}^{-1} \right)_k \]  

\[ - \left[ \left( \hat{R}^{-1} \right)_k X_k \right]^T \left( \hat{R}^{-1} \right)_k \left( \hat{R}^{-1} \right)_k^T X_k \]  

\[ = \frac{1}{1 - \alpha_k} \left[ \hat{X}_k^T \left( \hat{R}^{-1} \right)_k X_k \right] - \frac{\left( \hat{R}^{-1} \right)_k X_k \left( \hat{R}^{-1} \right)_k X_k \right]^T \]  

\[ + \frac{\left( \hat{R}^{-1} \right)_k X_k \left( \hat{R}^{-1} \right)_k X_k \right]^T \]  

\[ = \frac{1}{1 - \alpha_k} \left[ \left( \hat{R}^{-1} \right)_k X_k \right]^2 - \frac{\left( \hat{R}^{-1} \right)_k X_k \left( \hat{R}^{-1} \right)_k X_k \right]^T \]  

\[ + \frac{\left( \hat{R}^{-1} \right)_k X_k \left( \hat{R}^{-1} \right)_k X_k \right]^T \]  

\[ = \frac{1}{1 - \alpha_k} \left[ \left( \hat{R}^{-1} \right)_k X_k \right]^2 - \frac{\left( \hat{R}^{-1} \right)_k X_k \left( \hat{R}^{-1} \right)_k X_k \right]^T \]  

\[ + \frac{\left( \hat{R}^{-1} \right)_k X_k \left( \hat{R}^{-1} \right)_k X_k \right]^T \]
IV. SIMULATION RESULTS

A comparison of the variable step size and the fixed step size LMS/Newton algorithms was made for a system identification application. The chosen unknown system was a seventh-order lowpass filter. The input signal was a white noise with unit variance.

The number of iterations needed to achieve an MSE smaller than $10^{-10}$, for the identification of the filters mentioned above, was 79. The estimate of the mean square error at each instant of time was obtained by averaging the square error over 25 independent computer runs. The value of $b$ in the present example was $b = 1$.

In another experiment, the LMS/Newton algorithm with fixed $\alpha$ was applied to identify the seventh order filter, for a number of different values of $\alpha$. As can be seen from Table I the best results were achieved with $\alpha = 0.08$. The same problem was solved with the variable step algorithm, and as shown in Fig. 2 the average value of $\alpha_b$, taken from the last 100 samples, is very close to the best fixed $\alpha$ found experimentally.

It can be concluded that the variable step algorithm proposed, save the designer from choosing an appropriate value for $\alpha$. This guarantees fast convergence without any special attention to the properties of the input signal correlation matrix.

Faster convergence is achieved when $b$ is close to its lower bound since in this case $b \approx 0.5$ that is the optimum value for the ideal Newton algorithm. However, this value of $b$ can make the LMS/Newton diverge. In Table II, is illustrated the number of iterations required to identify the seventh order plant used in the previous examples. As can be noted values of $b$ closer to 0.5 yield faster convergence. In general, good results are obtained by choosing $b$ in the range $0.5 \leq b \leq 1$.

V. REFERENCES


Table I — Identification of a 7th order plant

<table>
<thead>
<tr>
<th>FIXED CONVERGENCE FACTOR $\alpha$</th>
<th>NUMBER OF ITERATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.04</td>
<td>124</td>
</tr>
<tr>
<td>0.06</td>
<td>97</td>
</tr>
<tr>
<td>0.08</td>
<td>88</td>
</tr>
<tr>
<td>0.10</td>
<td>92</td>
</tr>
<tr>
<td>0.12</td>
<td>106</td>
</tr>
<tr>
<td>0.14</td>
<td>120</td>
</tr>
</tbody>
</table>

Table II — Rate of convergence as a function of $b$ parameter

<table>
<thead>
<tr>
<th>$b$</th>
<th>NUMBER OF ITERATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>2/3</td>
<td>60</td>
</tr>
<tr>
<td>1</td>
<td>79</td>
</tr>
<tr>
<td>3/2</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>112</td>
</tr>
</tbody>
</table>

Fig. 1 — Adaptive FIR Filter

Fig. 2 — Variable Step Size Behavior.