MULTIPLE ENVIRONMENT OPTIMAL UPDATE PROFILING FOR STEEPEST DESCENT ALGORITHMS

Mile Milisavljević

Cicada Semiconductor Corporation
811 Barton Springs Road, Ste. 550
Austin, Texas 78704
mm@cicada-semi.com

ABSTRACT

In this paper, the methods for use of prior information about multiple operating environments, in improving adaptive filter convergence properties are discussed. More concretely, the gain selection, profiling and scheduling in steepest descent algorithms are treated in detail. Work presented in this paper is an extension of [1]. Two flavors of optimization are discussed: average descent rate optimization and maximization of the minimum descent rate. It is demonstrated, just as in the case of single channel optimization, with no additional complexity a substantial increase of convergence rate of steepest descent algorithms can be achieved. Finally, performance of the method is analyzed on the adaptive linear equalizer design for local area networks.

1. INTRODUCTION

In this section steepest descent adaptive filtering and basic single channel update profiling methods are described.

1.1. Steepest Descent Algorithms

Due to their low implementational cost and good numerical properties, steepest descent techniques play important role in modern signal processing applications. A typical application environment for steepest descent techniques in NC and EC is given in Figure 3.

The convergence of the iterative algorithm is governed by the difference equation:

$$\begin{aligned} \mathbf{v}_{n+1} &= (\mathbf{I} - \mu \mathbf{R}_{xx})\mathbf{v}_n \\ \mathbf{v}_n &= \mathbf{f}_n - \mathbf{f}^* \end{aligned} \tag{1}$$

where $\mathbf{f}$ represents the filter coefficient vector, $\mathbf{v}$ represents the filter coefficient error, and $\mathbf{R}_{xx}$ denotes the autocorrelation matrix of the signal . Stability of the SD adaptation can be derived from relationship (1) which indicates that in order to ensure the stability of the algorithm one needs to choose the adaptation step to be within bounds: $0 < \mu < \frac{2}{\lambda_{max}(\mathbf{R}_x)}$. Larger values of the adaptation step lead to faster convergence, but increase the residual error and may potentially render the algorithm unstable. In practical applications, such problems are usually resolved with adaptive change of the adaptation step size [2], or with a conservative choice of fixed step size[3].

Interestingly the convergence speed of the steepest descent algorithm is governed by the smallest eigenvalue of the correlation matrix [8] via:

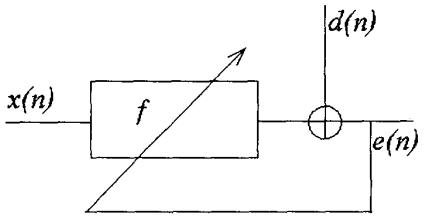

Fig. 1. Common adaptive system adapting to minimize the residual error.

$$\tau_n \approx \frac{1}{\mu \lambda_n}, \tag{2}$$

where, τ_n is the time constant corresponding to the eigenvalue λ_n. Equation (2) indicates that while the adaptation step-size is limited by the reciprocal of the largest eigenvalue, the smallest eigenvalue is the one that governs the convergence of the slowest mode. The relationships (1) and (2) expose the fundamental problem of application of steepest descent procedures to problems with large eigenvalue disparity.

1.2. Update Profiling

A practically more appealing version of steepest descent procedure is the least-mean-square (LMS) algorithm. In the LMS algorithm the gradient is substituted by its instantaneous estimate:

$$\nabla = \varepsilon_k \mathbf{x}_k = \mathbf{r}_{dx,\,k} - \mathbf{x}_k \mathbf{x}_k' \mathbf{f} \approx (d_k - \mathbf{x}_k' \mathbf{f})\mathbf{x}_k \tag{3}$$

Inclusion of the 3 in the gradient update yields:

$$\mathbf{f}_{n+1} = \mathbf{f}_n + \mu \varepsilon_n \mathbf{x_n}. \tag{4}$$

Definition: *Graded Update Gains* (or Graded Updates) refers to the application of different gain to every coefficient (tap) in the vector implementation of the steepest descent based algorithm.

0-7803-7041-4/01/$10.00 ©2001 IEEE

In graded update version therefore, every coefficient of the filter $\mathbf{f}$ has a different update rate μ_k. Thus (5) changes to:

$$\mathbf{f}_{n+1} = \mathbf{f}_n + \mathbf{M}\varepsilon_n \mathbf{x_n}. \quad (5)$$

where $\mathbf{M}$ is the diagonal matrix $\mathbf{M} = \mathrm{diag}(\mu_1, \mu_2, \ldots, \mu_N)$, and the error evolution equation (1) becomes:

$$\mathbf{v}_{n+1} = (\mathbf{I} - \mathbf{M}\mathbf{R}_{xx})\mathbf{v}_n \quad (6)$$

Most of authors dealing with this problem so far [5, 6] concentrated on the expected value of filter coefficients and adapted coefficients of larger magnitude with larger update gains. In this paper it is shown that contrary to the popular belief, optimal graded update gains may **only coincidentally** be connected to the expected value of coefficients.

The basic idea behind the solution of the graded updates problem is minimization of the expected error variance at every iteration. This optimization procedure yields (possibly time varying) set of gains $\mathbf{M}$ which allow fastest descent down the expected quadratic bowl.

In case when initial error statistics is unknown, or disregarded, the designer can then adopt the maximum entropy approach and assume white statistics of initial error. Let $\mathbf{Q}$ be the matrix that diagonalizes (6):

$$\tilde{\mathbf{v}}_{n+1} = \mathbf{Q}\mathbf{v}_{n+1} = \mathbf{Q}\left(\mathbf{I} - \mathbf{M}\mathbf{R}_{xx}\right)\mathbf{Q}'\mathbf{Q}\mathbf{v}_n = \mathbf{\Lambda}\tilde{\mathbf{v}}_n \quad (7)$$

Proposition 1: If $\mathbf{x}$ is white (i.e. $E\{\mathbf{x}\mathbf{x}'\} = \sigma^2\mathbf{I}$, and $\mathbf{Q}$ is an orthogonal matrix, then $\tilde{\mathbf{x}} = \mathbf{Q}\mathbf{x}$ is also white.
Proof: $E\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}'\} = \mathbf{Q}E\{\mathbf{x}\mathbf{x}'\}\mathbf{Q}' = \sigma^2\mathbf{Q}\mathbf{Q}' = \sigma^2\mathbf{I}$, QED.

Via Proposition 1, if $\mathbf{v}$ is assumed to be white, then $\tilde{\mathbf{v}}$ is also white. Expected value of the norm of $\tilde{\mathbf{v}}_{n+1}$ is then minimized when sum of squared eigenvalues (on the diagonal of) $\mathbf{\Lambda}$ is minimized, since:

$$\begin{aligned} E\{\tilde{\mathbf{v}}'_{n+1}\tilde{\mathbf{v}}_{n+1}\} = \; & E\{\mathbf{v}'\mathbf{\Lambda}^2\mathbf{v}\} = \\ & \textstyle\sum_{i=1}^{N} \lambda_i^2 E\{\tilde{v}_n^2(i)\} = \\ & \sigma^2_{\tilde{\mathbf{v}}_n} \textstyle\sum_{i=1}^{N} \lambda_i^2 \end{aligned} \quad (8)$$

The solution to this problem corresponds to the optimal $\mathbf{M}$ which is the diagonal approximation to the inverse of the $\mathbf{R}_{xx}$ in Frobenious norm:

$$\mathbf{M} = \mathrm{argmin}\|\mathbf{I} - \mathbf{M}\mathbf{R}_{xx}\|_F \quad (9)$$

An alternative solution to the problem can be reached if the initial error statistics is known. To derive the optimal graded update gains matrix $\mathbf{M}$, from (6) first compute the dynamics of the filter error norm $\mathbf{v}'_n\mathbf{v}_n$:

$$\mathbf{v}'_{n+1}\mathbf{v}_{n+1} = \mathbf{v}'_{n+1}\left(\mathbf{I} - \mathbf{M}\mathbf{R}_{xx}\right)'\left(\mathbf{I} - \mathbf{M}\mathbf{R}_{xx}\right)\mathbf{v}_{n+1} \quad (10)$$

In [1] it was shown that:

$$\begin{aligned} \sigma_v^2[n+1] \; &= \; E\{\mathbf{v}'_{n+1}\mathbf{v}_{n+1}\} \\ &= \; \mathrm{Tr}\left\{\mathbf{R}_{vv}\right\} - \\ &\quad \mathrm{Tr}\left\{\left(2\mathbf{R}'_{xx}\mathbf{M} - \mathbf{R}_{xx}\mathbf{R}'_{xx}\mathbf{M}\mathbf{M}\right)\mathbf{R}_{vv}[n]\right\} \end{aligned} \quad (11)$$

Noting that $\mathrm{Tr}\left\{\mathbf{R}_{vv}\right\} = E\{\mathbf{v}'_{n+1}\mathbf{v}_{n+1}\} = \sigma_v^2[n]$, we form the problem of finding $\mathbf{M}$ as an optimization problem:

$$\max_{\mathbf{M}\ \mathrm{diagonal}} \|\sigma_v^2[n] - \sigma_v^2[n+1]\| \quad (12)$$

Maximization of the difference between two subsequent errors ensures the maximization of the convergence rate of the algorithm. Cost function in (12) can be rewritten:

$$\min_{\mathbf{M}\ \mathrm{diagonal}} \mathrm{Tr}\left\{\mathbf{R}_{vv}[n]\mathbf{R}_{xx}\mathbf{R}'_{xx}\mathbf{M}\mathbf{M} - 2\mathbf{R}_{vv}[n]\mathbf{R}'_{xx}\mathbf{M}\right\} \quad (13)$$

Solution of this optimization problem is can be obtained by following optimization strategy [1, Proposition 3]: Consider the following optimization strategy for problem in (12): first find the optimal solution $\mathbf{M} = \mathbf{R}_{xx}^{-1}$, and then find the matrix $\mathbf{M}$ closest to the $\mathbf{R}_{xx}^{-1}$ in Frobenious sense. Such matrix will be the optimal constrained solution of problem in (12).

In other words, the autocorrelation matrix $\mathbf{R}_{xx}$ needs to be inverted, and its diagonal will represent the optimal update grading profile.

The fundamental difference between two solutions is in the optimization target. Optimization criteria in 8 minimizes sum of squared eigenvalues of matrix product $\mathbf{M}\mathbf{R}_{xx}$, while optimization criteria in 12 minimizes the maximum eigenvalue of the matrix product $\mathbf{M}\mathbf{R}_{xx}$. Clearly, from the speed of convergence perspective later optimization method is superior.

It is important to note, that above suggested solution is a close cousin of the Newton's algorithm based descent strategies [3]:

$$\mathbf{f}_{n+1} = \mathbf{f}_n + \mu\mathbf{R}_{xx}^{-1}\varepsilon_n\mathbf{x}_n \quad (14)$$

While full Newton's algorithm based descent strategies completely decouple and *equalize* rates of descent due to different modes, limitations to the diagonal matrix would achieve this only partially. Level of achieved benefit depends on the level of the input signal correlation.

2. MULTIPLE OPERATING ENVIRONMENT OPTIMAL UPDATE PROFILES

Main reason for using adaptive filters is the unknown and/or dynamic nature of the operating environment. In many cases, though, a designers may have a rough idea of operating environments adaptive systems may face. A good example of such case is a set of standard loops for wireline communications. When operating environment space can be characterized by representative samples of operating environments it is of interest to optimize the convergence performance of the steepest descent algorithms over such characterization. In this section two methods of multiple operating environment optimization: average convergence rate optimization and maximization of slowest convergence rate.

2.1. Average Convergence Rate Maximization

Let operating environment models be indexed from 1 to L. Furthermore, let them be characterized by the signal autocorrelation matrices $\mathbf{R}_{xx}^{(k)}$. The set of error vector $\mathbf{v}_n^{(k)}$ corresponding to each case of steepest descent application in environment k is defined as in (1) and (6):

$$\mathbf{v}_{n+1}^{(k)} = (\mathbf{I} - \mathbf{M}\mathbf{R}_{xx}^{(k)})\mathbf{v}_n^{(k)} \quad (15)$$

A super-vector of adaptation error can then be created to include all environments:

$$\mathbf{v}_{n+1} = \begin{bmatrix} \mathbf{v}_{n+1}^{(1)} \\ \mathbf{v}_{n+1}^{(2)} \\ \vdots \\ \mathbf{v}_{n+1}^{(L)} \end{bmatrix} = \mathbf{I} - \mathbf{R_M}\mathbf{v}_n = \\ = \mathbf{I} - \begin{bmatrix} \mathbf{M}\mathbf{R}_{xx}^{(1)} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{M}\mathbf{R}_{xx}^{(L)} \end{bmatrix} \begin{bmatrix} \mathbf{v}_n^{(1)} \\ \mathbf{v}_n^{(2)} \\ \vdots \\ \mathbf{v}_n^{(L)} \end{bmatrix} \quad (16)$$

where $\mathbf{R_M}$ represents the composite matrix with block diagonal sub-matrices $\mathbf{M}\mathbf{R}_{xx}^{(k)}$. The problem in (16) is similar to a single channel optimal profile problem. Solution which minimizes sum of squared eigenvalues can be computed by choosing $\mathbf{M}$ which gives minimizes $\|\mathbf{I} - \mathbf{R_M}\|_F$. Solution can be computed in the closed form by differentiation of $\|\mathbf{I} - \mathbf{R_M}\|_F$ with respect to diagonal entries m_j of matrix $\mathbf{M}$:

$$m_j = \frac{\sum_{k=1}^{L} r_{jj}^{(k)}}{\sum_{k=1}^{L}\sum_{i=1}^{N} r_{ij}^{(k)2}} \quad (17)$$

where $r_{ij}^{(k)}$ represents (i,j) element of $\mathbf{R}_{xx}^{(k)}$, and N is the dimensionality of $\mathbf{M}$.

2.2. Maximization of the Slowest Convergence Rate

Using the super-vector notation of (16) and following the method of maximization of the error vector norm descent. Having in mind the diagonal structure of $\mathbf{MR}$ and $\mathbf{M}$, it is possible to find a permutation matrix $\mathbf{P}$ such that:

$$\tilde{\mathbf{v}}_{n+1} = \mathbf{P}\mathbf{v}_{n+1} = \mathbf{P}(\mathbf{I} - \mathbf{R_M})\mathbf{P}'\mathbf{P}\mathbf{v}_n = (\mathbf{I} - \tilde{\mathbf{M}}\tilde{\mathbf{R}})\tilde{\mathbf{v}}_n \quad (18)$$

where $\tilde{\mathbf{M}}$ is a block constant diagonal matrix (i.e. it has first N diagonal elements equal to m_1, second N diagonal entries equal to m_2, etc.) and $\tilde{\mathbf{R}} = \mathbf{P}\mathbf{R}\mathbf{P}'$. Then, using the fact that the trace of a scalar is equal to the scalar and linearity of the expectation and trace operators derivation follows similarly as in Section 2:

$$\begin{aligned} \sigma_{\tilde{v}}^2[n+1] &= E\{\tilde{\mathbf{v}}_{n+1}'\tilde{\mathbf{v}}_{n+1}\} \\ &= E\left\{\mathrm{Tr}\left\{\tilde{\mathbf{v}}_n'\left(\mathbf{I} - \tilde{\mathbf{M}}\tilde{\mathbf{R}}\right)'\left(\mathbf{I} - \tilde{\mathbf{M}}\tilde{\mathbf{R}}\right)\tilde{\mathbf{v}}_n\right\}\right\} \\ &= \mathrm{Tr}\left\{\left(\mathbf{I} - \tilde{\mathbf{M}}\tilde{\mathbf{R}}\right)'\left(\mathbf{I} - \tilde{\mathbf{M}}\tilde{\mathbf{R}}\right)\mathbf{R}_{\tilde{v}\tilde{v}}[n]\right\} \\ &= \mathrm{Tr}\left\{\mathbf{R}_{\tilde{v}\tilde{v}}\right\}[n] - \\ &\quad \mathrm{Tr}\left\{\left(2\tilde{\mathbf{R}}'\mathbf{M} - \tilde{\mathbf{R}}\tilde{\mathbf{R}}'\tilde{\mathbf{M}}\tilde{\mathbf{M}}\right)\mathbf{R}_{\tilde{v}\tilde{v}}[n]\right\} \end{aligned} \quad (19)$$

Maximization of the difference between two subsequent errors ensures the maximization of the convergence of the slowest mode over all operating environments. Cost function can then be rewritten as:

$$\min_{\substack{\mathbf{M}\text{ diagonal} \\ \text{and block}-\text{constant}}} \mathrm{Tr}\left\{\mathbf{R}_{\tilde{v}\tilde{v}}[n]\tilde{\mathbf{R}}\tilde{\mathbf{R}}'\tilde{\mathbf{M}}\tilde{\mathbf{M}} - 2\mathbf{R}_{\tilde{v}\tilde{v}}[n]\tilde{\mathbf{R}}'\tilde{\mathbf{M}}\right\} \quad (20)$$

Just as in Section 2, and [1, Proposition 3], the solution of this problem is matrix $\mathbf{M}$ which most closely approximates the inverse of $\tilde{\mathbf{R}}$ and satisfies the block-constant and diagonality constraints. Given the convexity of constraints, following optimization strategy yields optimal grading profile: Inverse of $\tilde{\mathbf{R}}$ is computed, and it's diagonal is averaged in sections of N elements. Thus, a projection of $\tilde{\mathbf{R}}^{-1}$ on the constraint space is produced. Matrix $\mathbf{M}$ is therefore made up of diagonal elements:

$$m_i = \frac{1}{N}\sum_{j=1}^{N} \tilde{r}_{(i-1)N+j} \quad (21)$$

where $\tilde{r}_k$ is k^{th} diagonal element of $\tilde{\mathbf{R}}$.

3. EXAMPLES AND CONCLUSIONS

As an example of update profile use, consider equalization problem for local area network loops. Let the environment model space be represented by three channels: 100m Category 5 (CAT5) loop, 50m CAT5 loop, and 25m CAT5 loop. Channel responses normalized to unit energy (thus mimicking commonly used adaptive gain control) and shifted for delay are shown in Figure 2. Different channel spectral characteristics cause differences in autocorrelation matrices, thus warranting a conservative adaptive filter design. Matrix with largest eigenvalue spread then determines the update rate of classical steepest descent algorithms. In this case, 100m loop would determine the update rates. For brevity we demonstrate only performance of algorithm for maximization of the convergence rate of the slowest mode. Figure 3 shows the update grading profiles of 5 tap equalizer computed for each loop separately (solid = 100m loop, dashed = 50m loop, and dotted = 25m loop) as well as the joint grading profile (dash-dotted). Figure 4 shows the expected evolution distance in dB between computed filter coefficients f and optimal filter coefficients f^* for considered channels. Maximum update rate guaranteeing stability of over all channels is used for both graded and non-graded updates. Obviously for the loop with largest eigenvalue disparity the filter is slowest to converge to the optimal solution. Multiple channel grading profile in this case has largest impact on the convergence and increases convergence rate considerably. For 50m loop, grading profile increases convergence rate a bit, and for 25m loop it decreases convergence rate. However, 100m loop being the limiting factor, overall performance is significantly improved. It is important to note that (in this case) equalizer mean square error performance depends not only on distance between f and f^* but also on sensitivity of the error (i.e.steepness of the quadratic bowl around the optimal solution). Hence, this method dramatically improves adaptation properties in precision sensitive cases.

4. REFERENCES

[1] Milisavljević, M., "Optimal Update Profiling for Steepest Descent Algorithms", Proceedings of DSP2000 Workshop, Hunt, TX, October 2000.

[2] Bishop, C.M., "Neural Networks for Pattern Recognition", Clarendon Press - Oxford 1995.

[3] Clarkson, P. M., "Optimal and Adaptive Signal Processing", CRC Press 1993.

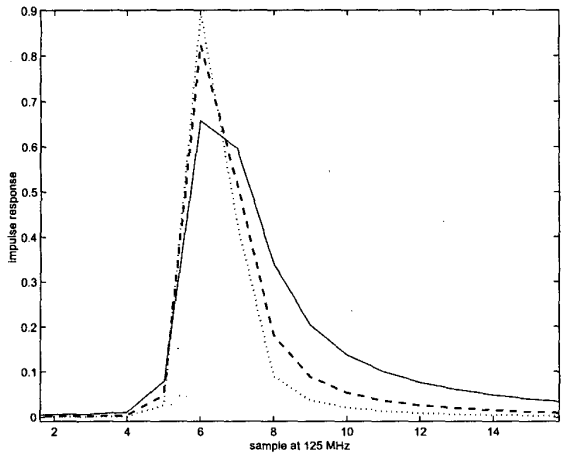

Fig. 2. Channel impulse responses under consideration under consideration: solid = 100m CAT5 loop, dashed = 50m CAT5 loop, and dotted =25m CAT5 loop.

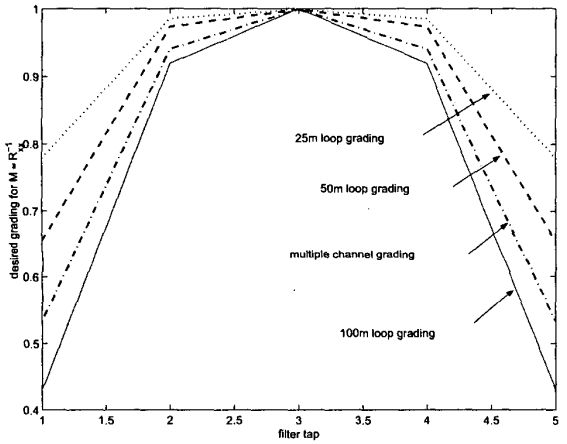

Fig. 3. Computed grading profiles for 3 channels, and slowest mode optimizing profile.

[4] Golub, G. and Van Loan, C., "Matrix Computations", Johns Hopkins 1996.

[5] Makino, S., Kaneda, Y., Koizumi, N., "Exponentially Weighted Stepsize NLMS Adaptive Filter Based on Statistics of a Room Impulse Response", IEEE Transactions on Speech and Audio Processing, Vol. 1, No. 1, January 1993.

[6] McCaslin, S. and Van Bavel, N., "Effects of Quasi-Periodic Training Signals on the Performance of Acoustic Echo Cancellers", Annales des Telecommunications, Vol. 49, No. 7-8, pp. 380-385, July 1994.

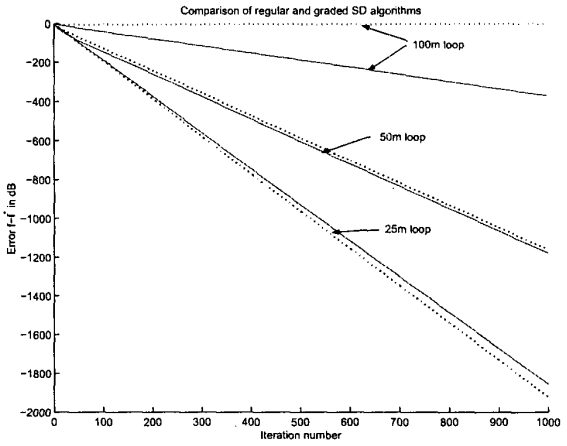

Fig. 4. Convergence towards optimal solution with updates graded for multi-channel use. Solid lines = graded updates, dotted lines = scalar update.

[7] Rupp, M. "Bursting the LMS Algorithm", IEEE Transactions on Signal Processing, Vol. 43, No. 10, October 1995.

[8] Widrow, B. and Stearns, S.D., "Adaptive Signal Processing", Engelwood Cliffs, N.J. Prentice Hall 1985.