Reprinted from IEEE CONFERENCE RECORD OF THE FOURTEENTH
ASILOMAR CONFERENCE ON CIRCUITS, SYSTEMS & COMPUTERS, November 1980

ON THE SPEED OF CONVERGENCE OF ADAPTIVE ALGORITHMS

by B. Widrow

Department of Electrical Engineering
Stanford University, Stanford, California

Two forms of the LMS adaptive algorithm will be discussed here, the "usual" algorithm based on the method of steepest descent [1-5], and an idealized algorithm based on Newton's method. These algorithms will be considered from the points of view of: (a) rate of convergence, (b) efficiency of statistical performance.

Speeding up a given adaptive process generally requires that the adaptive parameters (weights, etc.) take values based on averaging over less input data. The result is increased parameter noise and reduced average system performance. When using a specific algorithm, there is generally a tradeoff between speed of convergence and average statistical performance.

Two algorithms may be compared with each other when applied to the same adaptation task by adjusting their rates of convergence to cause the same effective parameter noise. As such, the more efficient algorithm converges faster. Effective parameter noise is that attribute of the noise that causes loss in system performance.

Referring to the literature [1-5], the steepest descent version of the LMS algorithm is

$$W_{j+1} = W_j + 2\mu\varepsilon_j X_j \quad . \tag{1}$$

The p^{th} mode of the mean square error learning curve has a time constant given by

$$\tau_{p_{mse}} = \frac{1}{4\mu\lambda_p} \quad . \tag{2}$$

It is seen that increasing the convergence factor μ speeds up the adaptive process by causing the adaptive time constants to be reduced. However, to insure stability in the mean, μ must be kept within the bounds

$$\frac{1}{\lambda_{max}} > \mu > 0 \quad , \tag{3}$$

where λ_{max} is the largest eigenvalue of the input correlation matrix R. After adaptive transients die out, noise in the weights causes, on the average, an increase in mean square error over the theoretical minimum mean square error. The misadjustment M has been defined [1-5] as the dimensionless ratio of the average excess mean square error to the minimum mean square error. For the steepest descent LMS algorithm,

$$M = \mu \text{ trace } R = \frac{n}{4}\left(\frac{1}{\tau_{p_{mse}}}\right)_{ave} \quad . \tag{4}$$

Increasing μ speeds up the adaptive process but increases the misadjustment.

A "Newton's method" version of the LMS algorithm premultiplies the instantaneous gradient estimate $2\varepsilon_j X_j$ by the inverse of R. The algorithm is

$$W_{j+1} = W_j + 2\mu\lambda_{ave}\varepsilon_j R^{-1} X_j \quad . \tag{5}$$

The scaling constant λ_{ave} (the average of the eigenvalues) has been included for convenience. It can be shown that premultiplication by R^{-1} causes each adaptive step to be taken not along the maximum gradient but instead in the direction toward the minimum of the mean square error performance surface, toward the bottom of a quadratic bowl. The effect is very much like application of steepest descent when all eigenvalues are equal. The eccentricity of the performance function is eliminated by Newton's method. The only drawback is that Newton's method as specified by (5) requires R^{-1} which is generally not available. The unavailability of R^{-1} is often the reason for using an adaptive process in the first place. An attempt to perform an algorithm like equation (5), only using an R^{-1} estimated from input data, has been reported by Griffiths and Mantey [6]. We shall focus our attention on equation (5), realizing that such an algorithm, a true Newton's method version of LMS, is a mathematical idealization. It can be shown to have the following properties. Instead of there being a number of time constants of the mean square error learning curve equal to the number of weights (as with conventional LMS), there is a single time constant given by

$$\tau_{mse} = \frac{1}{4\mu\ \lambda_{ave}} \quad . \tag{6}$$

The misadjustment of algorithm (5) is given by

$$M = \mu \text{ trace } R = \frac{n}{4}\left(\frac{1}{\tau_{mse}}\right) \quad . \tag{7}$$

The bounds on μ for convergence in the mean are

$$\frac{1}{\lambda_{ave}} > \mu > 0 \quad . \tag{8}$$

Comparing the two algorithms, we make their μ-values equal in order to have equal misadjustments. Immediately we see that the stable range of μ for steepest descent is smaller than for Newton's method when there is eigenvalue disparity. Since these algorithms are generally operated with small μ to maintain small M, this is not necessarily disadvantageous for steepest descent. However, when the eigenvalue spread is extreme, steepest descent may be forced to operate with a very small value of μ in order to maintain stability. Under such circumstances, steepest descent would be stability bound rather than misadjustment bound.

With equal μ, it is interesting to compare the Newton's method time constant (6) with the steepest descent time constants (2). It is clear that some of the steepest descent convergence modes are going to be faster while some are going to be slower than the Newton's method mode. Initial conditions will determine the relative strengths of the various steepest descent modes. If we compare areas under the learning curves in order to compare "learning times" of single mode exponential curves with multimode curves (as is done in Fig. 1), it can be shown that when misadjustment bound, the learning time of steepest descent averaged over random initial conditions is identical to the learning time of Newton's method. However, one should realize that the worst case learning time for steepest descent will be worse than that for Newton's method by a factor of $\lambda_{max}/\lambda_{ave}$.

The behavior of the steepest descent LMS algorithm has been analyzed in detail in [4] with a simple form of nonstationary input that results in the quadratic mean square error function undergoing a random vector displacement. The motion of the bottom of the bowl is first order Markov. Misadjustment results both from noise in the weights and from the weights dynamically lagging behind the bottom of the moving mean square error bowl. It is shown that the total misadjustment is minimized when the rate of adaptation is adjusted (by choice of μ) so that both components of misadjustment are equal. A similar analysis has been made for LMS Newton, and it has been found that the value of μ that optimizes steepest descent also optimizes Newton's method and that both algorithms yield the same misadjustment for the same μ. The conclusion is that if the steepest descent algorithm is misadjustment bound rather than stability bound, steepest descent gives identical performance in a statistical sense to Newton's method with simple nonstationary inputs.

The Newton's method version of the LMS algorithm is about as efficient as an algorithm can be, from the standpoint of statistical performance. For a given number of weights and for a given level of misadjustment, the number of data samples seen and consumed in the convergence process of LMS Newton is about as small as nature will permit. Justification for this comes from study of adaptive behavior when learning with a finite number of data samples.

It is shown in Appendix A of reference [4] that when training an n-weight adaptive system with N independent data vectors, the expected misadjustment is

$$M = \frac{n}{N} = \frac{\text{number of weights}}{\text{number of training samples}} \quad . \tag{9}$$

This result was first reported in [1]. It is independent of algorithm, as long as the algorithm is least squares. A few simplifying assumptions were made to arrive at such an elegant and simple result. This formula has been tested extensively by computer simulation and has been found to be quite accurate for misadjustments of 25% or less.

Misadjustment formula (9) may be compared with that of LMS Newton (7). The comparison cannot be exact however, because (9) applies to learning with finite data while (7) applies to a steady flow learning process. Actually, (9) applies to steady flow learning with a uniform moving-average window while (7) applies to steady flow learning with an exponential moving window. Reconsider equation (7). It can be written as

$$M = \frac{n}{4\tau_{mse}} \quad . \tag{10}$$

One can read this as "misadjustment of LMS Newton equals the number of weights divided by the number of training samples," if one considers that an exponential process essentially settles within four time constants and that any input that has occurred more than four time constants ago would have negligible effect on the weights. We conclude that LMS Newton has the misadjustment of a fundamental least squares process. No adaptive process could have a lower misadjustment than (9) or (10). No more "information" (in the common English sense) can be squeezed from a given amount of data.

We now know that the Newton's method version of LMS is as fine and efficient as an adaptive algorithm can be. Unfortunately it cannot be implemented unless one perfectly knows R^{-1}. Without such knowledge, one can only approximate LMS Newton. In application to adaptive FIR digital filters, perhaps this or something approximating this is done by the adaptive lattice-filter algorithms that have appeared during the last half dozen years or so. It remains to be seen.

With extreme eigenvalue disparity, stability may be of limiting concern rather than misadjustment. To achieve maximum convergence speed with LMS steepest descent, μ would be set to $1/\lambda_{max}$, causing the slowest mode to have a time constant of $\lambda_{max}/4\lambda_{min}$ adaptations. Operating steepest descent at full speed, the misadjustment would be $M = \text{trace } R/\lambda_{max} \geq 1$. LMS Newton, doing the same job, could be pushed much faster. Maximum speed would be achieved by setting μ to half its upper stable limit, i.e. $\mu = 1/2\ \lambda_{ave}$, giving theoretical (no gradient noise) convergence in one iteration. As such, its misadjustment would be $M = n/2$. A 100 weight filter, for example, would have a misadjustment of 50.

In most engineering applications, a misadjustment of 1 would be considered very high. The

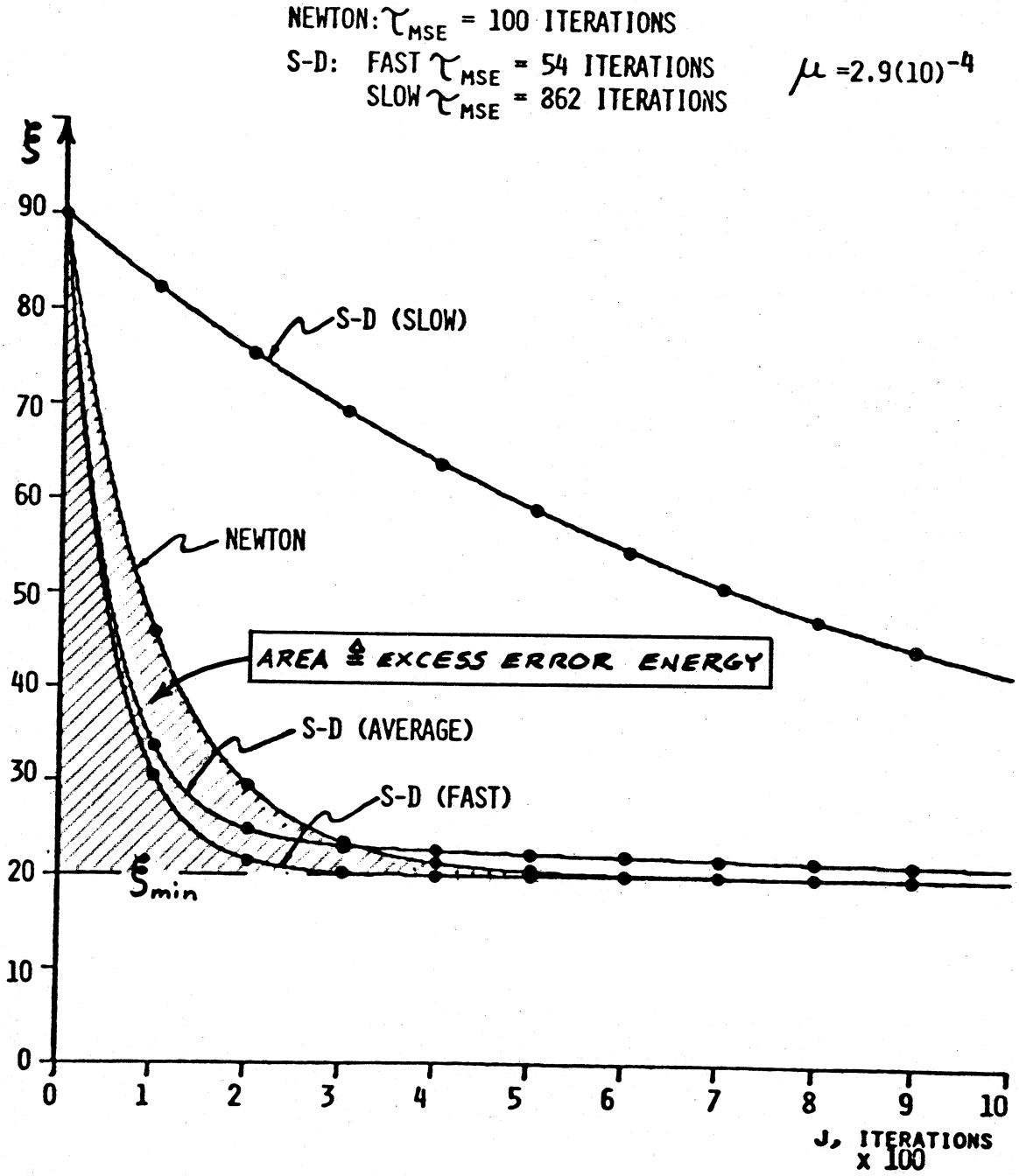

FIG. 1. LEARNING CURVES FOR NEWTON'S METHOD AND THE METHOD OF STEEPEST DESCENT (S-D)

adaptive solution then gives twice as much mean square error as the Wiener solution, i.e. 3 dB more mean square error. Only when the minimum mean square error of the Wiener solution is zero or very small would high misadjustment be acceptable. Input noise precludes this situation however.

In most engineering applications, a misadjustment of about 10% would be satisfactory. As such, neither LMS steepest descent nor LMS Newton would be pushed anywhere near to the brink of instability. Speed of convergence would then be misadjustment limited rather than stability limited, regardless of eigenvalue spread.

It is safe to say that the steepest descent version of the LMS algorithm is the simplest, most widely used, and most widely understood of all adaptive algorithms. With all eigenvalues equal, its performance is identical to that of LMS Newton. With eigenvalue disparity, its average speed of convergence and statistical efficiency and performance with nonstationary inputs is identical to that of LMS Newton, except that its worst case convergence rate is poorer.

When dealing with an input signal buried in noise, the eigenvalue spread is slight and one can be assured that LMS steepest descent will perform as well or better than any other algorithm. When the input is noise free or only slightly noisy and when the input signal is narrow-band with a highly peaked spectrum, eigenvalue disparity could be large or extreme. Under these circumstances, opportunities may or may not exist to better the performance of the steepest descent form of the LMS algorithm.

References

[1] B. Widrow and M. E. Hoff, "Adaptive Switching Circuits," in 1960 WESCON Conv. Rec., pt. 4, pp. 96-140.

[2] B. Widrow, P. Mantey, L. Griffiths, and B. Goode, "Adaptive Antenna Systems," Proc. IEEE, vol. 55, pp. 2143-2159, Dec. 1967.

[3] B. Widrow et al., "Adaptive Noise Cancelling: Principles and Applications," Proc. IEEE, vol. 63, pp. 1692-1716, Dec. 1975.

[4] B. Widrow, J. M. McCool, M. G. Larimore, and C. R. Johnson, Jr., "Stationary and Nonstationary Learning Characteristics of the LMS Adaptive Filter," Proc. IEEE, vol. 64, no. 8, pp. 1151-1162, August 1976.

[5] B. Widrow and J. M. McCool, "A Comparison of Adaptive Algorithms Based on the Methods of Steepest Descent and Random Search," IEEE Trans. on Antennas and Propagation, vol. AP-24, no. 5, pp. 615-637, Sept. 1976.

[6] L. J. Griffiths and P. E. Mantey, "Iterative Least-squares Algorithm for Signal Extraction," in Proc. Second Hawaii Int. Conf. System Sciences, Western Periodicals Co., pp. 767-770, 1969.