On the Speed of Training Networks with Correlated Features

Robert R.N. Bakker, Martin A. Kraaijveld, Robert P.W. Duin and Wouter F. Schmidt
Pattern Recognition Group
Department of Applied Physics
Delft University of Technology
P.O. Box 5046
2600 GA Delft
The Netherlands
e-mail: martin@ph.tn.tudelft.nl

Abstract — The learning speed of the adaptive linear combiner is determined by the condition number of the input correlation matrix of the training data. With known properties of such correlation matrices, it will be shown that increasing the dimensionality of the feature space of an adaptive linear combiner will never increase its learning speed. In fact, the learning speed will at best remain equal, but will deteriorate in most cases. Our result can be applied in adaptive learning problems that are time constrained, and has possibly implications for the training of multi-layer networks.

I. Introduction

An important problem in adaptive signal processing and pattern recognition applications, is the determination of the number of features (i.e. the dimensionality of the input vector) for a learning system. Obviously when the number of features is high much information is being processed, and it is to be expected that a high accuracy can be reached, e.g. in a mean square sense [1], or with respect to the classification performance [2]. A second issue that is especially relevant in time constrained applications, is how the *learning speed* is being influenced by the dimensionality of the input data. To be more specific, in this paper we will investigate how the learning speed of a class of iterative learning procedures is affected by adding an extra feature. This class consists of the steepest descent procedure and the LMS procedure for the Adaptive Linear Combiner, the ALC, see [1]. For these procedures the learning speed is determined by the condition number of the input correlation matrix of the data; i.e. the ratio of the smallest and largest eigenvalues of the input correlation matrix.

A result from linear algebra will be discussed, that proves that the learning speed of the adaptive linear combiner is decreased, when a new correlated feature is added. This is based on the fact that the largest eigenvalue generally becomes larger and the smallest eigenvalue generally becomes smaller, due to adding the new feature. Only when the new feature is not correlated with the old features and it is properly scaled, the learning speed will not be affected.

In the following paragraphs, we will review how the learning speed is influenced by the condition number of the input correlation matrix of the training data. Then, in section 3, it will be proven that the condition number decreases when a new correlated feature is added. Finally, the discussion and conclusions are presented in section 4.

2. The learning speed and the condition number for the ALC.

An adaptive linear network, like the ALC, is described by the following simple function:

$$y_t = X_t^T W_t = \sum_{i=1}^{n} x_{t,i} w_{t,i} \tag{1}$$

That is, where t is the time, the scalar output y_t is determined by the inner product of the (n- dimensional) input vector X_t and the weight vector W_t. During an iterative learning procedure, the weight vector W_t is adjusted such that the output y_t approximates a certain desired output as good as possible. For this learning procedure a set of pairs of input vectors X_t with corresponding desired output d_t is used. This learning set is supposed to represent the underlying phenomenon sufficiently well; i.e. the size of the learning set

This work was sponsored by the Dutch Government as a part of the SPIN / FLAIR-DIAC project, and by the Foundation of Computer Science in the Netherlands (SION) with financial support from the Dutch Organization for Scientific Research (NWO).

0-7803-0999-5/93/$03.00 ©1993 IEEE

should be large enough so that the learning set is statistically representative for the underlying problem.

As a measure for the quality of the approximation, a squared error criterion can be taken:

$$\mathrm{MSE} \equiv \xi_t = E\left[(y_t - d_t)^2\right] = E\left[\varepsilon_t^2\right] \tag{2}$$

In order to make the necessary definitions, we follow Widrow [1] and express this quadratic error criterion in terms of some statistics of the training data:

$$E\left[\varepsilon_t^2\right] = E\left[d_t^2\right] + W_t^T \mathbf{R} W_t - 2P^T W_t \tag{3}$$

Here **R** is the semi-positive definite symmetric input correlation matrix, defined as $E\left[X_t X_t^T\right]$ and P is the vector $E\left[d_t X_t\right]$. From (3) it is clear that the MSE is a quadratic form in W_t. The optimal weight vector W^*, i.e. the location where ξ has its minimal value ξ_{min}, is found by taking the derivative with respect to W. ξ_t can be shown to depend only on the distance between the actual and optimal weight vectors and the input correlation matrix:

$$\xi_t = \xi_{min} + \left(W_t - W^*\right)^T \mathbf{R}\left(W_t - W^*\right) \tag{4}$$

Now the coordinate system W can be translated to a system V, such that $\xi_{min} = \xi|_{W^*}$ is located at the origin, followed by a rotation to a coordinate system V' in which the V'-axes coincide with the principal axes of the paraboloid ξ-surface:

$$\begin{aligned} \xi_t &= \xi_{min} + \left(W_t - W^*\right)^T \mathbf{R}\left(W_t - W^*\right) \\ &= \xi_{min} + V_t^T \mathbf{R} V_t \\ &= \xi_{min} + V_t^T\left(Q \Lambda Q^T\right) V_t \\ &= \xi_{min} + V'^T_t \Lambda V'_t \end{aligned} \tag{5}$$

Due to this transformation, **R** is replaced by its eigenvalue matrix Λ. The eigenvectors Q_i ($1 \le i \le n$) of the input correlation matrix define the principal axes of the error surface and the eigenvalues λ_i ($1 \le i \le n$) give the second derivatives of the error surface ξ with respect to its principal axes.

In the case of steepest descent minimization of ξ, the weight vector is adjusted as follows:

$$W_{t+1} = W_t - \eta \nabla_t \tag{6}$$

where ∇_t is the gradient and η represents the step-size or learning rate. Due to this iterative learning procedure, the trajectory in weight space can be expressed as*:

$$W_t = W^* + (\mathbf{I} - 2\eta\mathbf{R})^t\left(W_0 - W^*\right) \tag{7}$$

Note that (7) gives the weight vector at any time t as a function of the weight vector W_0 at time zero. In terms of V' (7) can be formulated as:

$$V'_t = (\mathbf{I} - 2\eta\Lambda)^t V'_0 \tag{8}$$

from which it is apparent that for a stable learning behavior we need to choose:

$$0 < \eta < \frac{1}{\lambda_{max}} \tag{9}$$

Also, from (8) it is clear that the speed of the learning procedure is limited by the smallest eigenvalue λ_{min}, since the term with λ_{min} is the slowest converging term in (8).. Apparently, the condition number of **R**, defined as the ratio of the smallest and the largest eigenvalue, is a good measure for the speed of the iterative learning procedure. In the next section the effect on the learning speed and the condition number caused by adding an extra feature x_{n+1} will be discussed.

3. ADDING A (CORRELATED) FEATURE.

Adding a new feature to the existing set of n features, implies that the matrix **R** is bordered with an n-dimensional vector U and a new diagonal element σ^2.

$$\mathbf{R}^{n+1} = \begin{pmatrix} \left[\mathbf{R}^n\right] & U \\ U^T & \sigma^2 \end{pmatrix} \tag{10}$$

The question that should be answered now is how the eigenvalues of **R** are affected by this bordering operation? The answer to this question can be found in a number of textbooks on linear algebra, e.g. see [3 - 5]. It is based on the Courant-Fisher theorem and the Rayleigh theorem and states that bordering the matrix will not decrease the largest eigenvalue, and not increase the smallest eigenvalue of the matrix. The condition number of the matrix will therefore deteriorate, or at best remain equal.

* For the LMS learning procedure, or Widrow-Hoff rule [Duda 1973], the gradient is replaced by an unbiased estimate of the gradient. Equation (7) is then approximately valid.

Here, we will go one step further than the previous (qualitative) proofs, by not only showing that the condition number deteriorates, but we will also provide an implicit equation for the new eigenvalues in terms of the old eigenvalues.

A new eigenvector Q', associated with a λ', obeys the following relation:

$$\mathbf{R}^{n+1}Q' = \lambda' Q' \qquad (11)$$

Now Q' is written as combination of an n-dimensional vector Ω and a scalar β:

$$\begin{pmatrix} \mathbf{R}^n & U \\ U^T & \sigma^2 \end{pmatrix} Q' = \lambda' Q' \Leftrightarrow \begin{pmatrix} \mathbf{R}^n & U \\ U^T & \sigma^2 \end{pmatrix} \begin{pmatrix} \Omega \\ \beta \end{pmatrix} = \lambda' \begin{pmatrix} \Omega \\ \beta \end{pmatrix} \qquad (12)$$

This equality can be split into two parts. By writing Ω as a vector Ψ that is rotated by the matrix $\mathbf{Q}$, the first part can be expressed as a matrix equation of the form:

$$\left.\begin{array}{l} \mathbf{R}^n\Omega + \beta U = \lambda' \Omega \\ \Omega = \mathbf{Q}\Psi \end{array}\right\} \mathbf{R}^n\mathbf{Q}\Psi + \beta U = \lambda' \mathbf{Q}\Psi \Leftrightarrow$$

$$\mathbf{Q}\mathbf{\Lambda}\Psi + \beta U = \lambda' \mathbf{Q}\Psi \qquad (13)$$

Note that equation (5) and the property $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$ were used. The second equation that follows from (12) is the scalar equation:

$$\left.\begin{array}{l} U^T\Omega + \sigma^2\beta = \lambda' \beta \\ \Omega = \mathbf{Q}\Psi \end{array}\right\} U^T\mathbf{Q}\Psi = \beta\left(\lambda' - \sigma^2\right) \qquad (14)$$

If (13) is multiplied with $\mathbf{Q}^T$ on both sides of the equality sign, we find that:

$$\mathbf{\Lambda}\Psi + \beta\mathbf{Q}^T U = \lambda' \Psi \Leftrightarrow \Psi = \beta(\lambda' \mathbf{I} - \mathbf{\Lambda})^{-1}\mathbf{Q}^T U \qquad (15)$$

Substituting Ψ in (14) yields:

$$U^T\mathbf{Q}\left\{\beta(\lambda' \mathbf{I} - \mathbf{\Lambda})^{-1}\right\}\mathbf{Q}^T U = \beta\left(\lambda' - \sigma^2\right) \qquad (16)$$

Which can finally be rewritten as:

$$\sum_{i=1}^{n} \frac{\left(U^T Q_i\right)^2}{(\lambda' - \lambda_i)} = \left(\lambda' - \sigma^2\right) \qquad (17)$$

This is an implicit equation for the new eigenvalues λ' in terms of the old eigenvalues λ_i. It is easily verified that the derivative of the left part of equation (17) with respect to λ' is always negative, with asymptotes for $\lambda' = \lambda_i$. The new eigenvalues are found at the locations where the left part of (17) crosses the linear function ($\lambda' - \sigma^2$), see figure 1. From figure 1 it is also seen that there is a new eigenvalue left of each old eigenvalue, and one extra eigenvalue being larger than the largest original eigenvalue. This qualitative finding confirms the proofs that are based on the Courant-Fisher and Rayleigh theorems, e.g. see [3 - 5]. Note that the eigenvalues do not change when the new feature is not correlated with the previous features, i.e. when $U = (0, 0, ..., 0)^T$. Also note that the fact that $\mathbf{R}^{n+1}$ is a correlation matrix assures that all new eigenvalues are non-negative.

4. DISCUSSION AND CONCLUSIONS.

Two direct conclusions of the foregoing theory are, that the smallest eigenvalue will not increase nor will the largest eigenvalue decrease by adding an extra feature. The first conclusion implies that the rate of convergence of equation (8) becomes slower, and the second conclusion necessitates the choice of a smaller step-size η (in equation 9), thereby also constraining the learning speed. It is important to note that especially this second aspect restricts the learning speed in practice. In practical applications the smallest eigenvalues may have neglectable influence on the final error ξ, so the learning phase is generally terminated after some application dependent time interval. Small eigenvalues becoming smaller will probably not influence this time interval. The fact that the largest eigenvalue becomes larger is therefore of much more influence, since a smaller step-size constrains the speed of the total learning process.

It should also be noted that besides these fundamental restrictions on the learning speed, there is also the fact that an increase of the number of features requires more computations per step. This was not taken into account here, but may be a relevant factor in practical applications.

Another issue that needs to be addressed is that our results are in line with previous results that claim that the learning speed is increased by decorrelating and scaling of the data, e.g. see [6] and [7]. From (8) it is easily seen that the learning speed is maximum when all eigenvalues are equal, whereas our results show that only when the new feature is not correlated and properly scaled there is no decrease in learning speed.

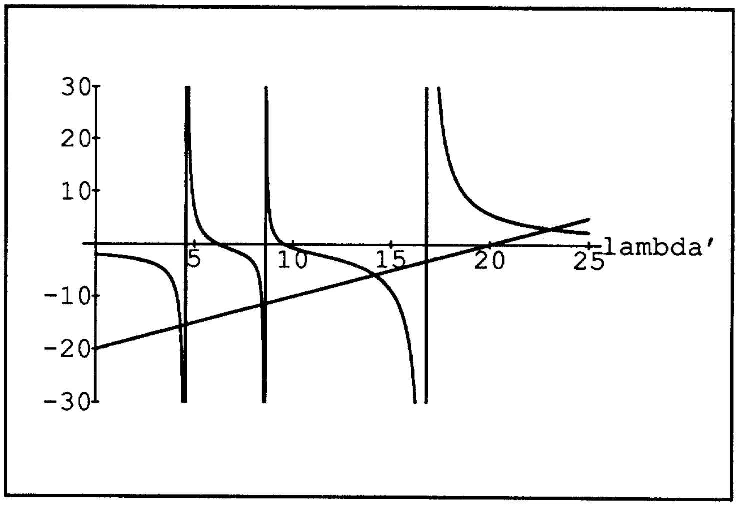

Figure 1: A numerical example of the theory. The matrix ((5, 1, 2), (1, 10, 3), (2, 3, 15)) with eigenvalues 4.58, 8.59 and 16.83 is bordered with the vector (1, 2, 3, 20)^T. The left side of equation (17) is the function with asymptotes at the original eigenvalues. The linear function (λ' - 20) corresponds to the right side of (17). The new eigenvalues are given by the locations where the two functions cross, i.e. at 4.57, 8.57, 14.42 and 22.45.

Finally, an important question is how our result generalizes towards more complex adaptive systems like multi-layer feed forward networks, e.g. see [8]. The problem with multi-layer feed forward networks is that the system is not linear, which has the effect of local minima in the error landscape. Obviously, where the error landscape can be locally approximated by a second order approximation, our result holds. Therefore, (as long as the approximation is valid) the learning will locally slow down. When the approximation is not valid anymore, the problem may emerge that the extension of the parameter space induces a new path to a close local minimum. This local minimum may then be reached much faster than some (local) minimum in the original space. One conclusion that does generalize towards non-linear systems, however, is that the learning rate should be taken smaller when the number of features is increased.

References

[1] Widrow, B., and Stearns, S.D., *Adaptive Signal Processing*, Prentice-Hall, Englewood Cliffs, New Jersey 1985.

[2] Jain, A.K., and Chandrasekaran, B., "Dimensionality and Sample Size Considerations in Pattern Recognition Practice", in: P.R. Krishnaiah and L.N. Kanal, editors, *Handbook of Statistics*, Vol. 2, North Holland Publishing Company, pp. 835-855, 1982.

[3] Lancaster, P., and Tismenetsky, M., *The Theory of Matrices*, second edition, Academic Press, 1985.

[4] Horn, R.A., and Johnson, C.R., *Matrix Analysis*, Cambridge University Press, Cambridge, 1988.

[5] Strang, G., *Linear Algebra and its Applications*, Harcourt Brace Jovanovich Publishers, San Diego, 1988.

[6] le Cun, Y., Kanter, I., and Solla, S.A., "Eigenvalues of Covariance Matrices: Applications to Neural-Network Learning", Physical Review Letters, Vol. 66, Nr. 18, May 6, 1991, pp. 2396 - 2399.

[7] Orfanidis, S.J., "Gram-Schmidt Neural Nets", Neural Computation, Vol. 2, pp. 116 - 126, 1990.

[8] Rumelhart, D.E., Hinton, G.E., and Williams, R.J., "Learning Internal Representations by Error Propagation", in: *"Parallel Distributed Processing: Explorations in the Microstructure of Cognition"*, Vol. 1, Rumelhart, D.E., and McClelland, J.L. (eds.), Cambridge, MA., MIT-Press, pp. 318-362.