Communicated by Erkki Oja

Temporal Evolution of Generalization during Learning in Linear Networks

Pierre Baldi
Jet Propulsion Laboratory and Division of Biology,
California Institute of Technology, Pasadena, CA 91125 USA

Yves Chauvin
Department of Psychology, Stanford University,
Stanford, CA 94305 USA
and
NET-ID, Inc., Menlo Park, CA 94025 USA

We study generalization in a simple framework of feedforward linear networks with n inputs and n outputs, trained from examples by gradient descent on the usual quadratic error function. We derive analytical results on the behavior of the validation function corresponding to the LMS error function calculated on a set of validation patterns. We show that the behavior of the validation function depends critically on the initial conditions and on the characteristics of the noise. Under certain simple assumptions, if the initial weights are sufficiently small, the validation function has a unique minimum corresponding to an optimal stopping time for training for which simple bounds can be calculated. There exists also situations where the validation function can have more complicated and somewhat unexpected behavior such as multiple local minima (at most n) of variable depth and long but finite plateau effects. Additional results and possible extensions are briefly discussed.

1 Introduction

Generalization properties of neural networks trained from examples seem fundamental to connectionist theories but also poorly understood. In practice, the question to be answered is how should one allocate limited resources and parameters, such as network size and architecture, initial conditions, training time, and available examples, to optimize generalization performance? One conventional approach is to consider the problem of learning as a surface fitting problem. Accordingly, neural networks should be very constrained, with a minimal number of parameters, to avoid the classical "overfitting" problem. In practice, however, not too much is known about overfitting, its nature, and its onset both

as a function of network parameters and training time. Furthermore, the conventional view has sometimes been challenged in light of simulation results and may need to be revised to some extent. It may be the case, for instance, that a suitable strategy consists rather in using networks with a few more parameters than the most constrained ones and training these slightly larger networks for shorter times, based on a careful monitoring of the evolution of the validation error during training and its minimization.

Partial interesting results on generalization have been obtained in recent years in terms of VC dimension and statistical mechanics (see, for instance, Baum and Haussler 1989; Tishby *et al.* 1989; and Sompolinsky *et al.* 1990). Most of these results, however, are static in the sense that they study generalization as a function of network architecture and number of examples. Here, we propose a different and complementary approach consisting in a detailed analysis of the temporal evolution of generalization in simple feedforward linear networks. This setting is not as restricted as it may seem because parametrically linear networks have been gaining popularity recently (e.g., radial basis functions or polynomial networks). Additional motivation for investigating these architectures can be found in Baldi and Hornik (1989, 1991). Even in this simple framework, the question is far from trivial. Thus we have restricted the problem even further: learning the identity map in a single layer feedforward linear network. With suitable assumptions on the noise, this problem turns out to be insightful and to yield analytical results that are relevant to what one observes in more complicated situations. With hindsight, it is rather remarkable that the complex phenomena related to generalization that are observed in simulations of nonlinear networks are already present in the linear case.

In Section 2, we define the framework and derive the basic equations first in the noiseless case and then in the case of noisy data. The basic point is to derive an expression for the validation function in terms of the statistical properties of the population and the training and validation samples. Section 3 contains the main results, which consist of an analysis of the landscape of the validation error as a function of training time. Simple simulation results are also presented and several interesting phenomena are described. The results are discussed and some possible extensions are briefly mentioned in the conclusion. Mathematical proofs are deferred to the Appendix.

2 Formal Setting

2.1 Noiseless Data. We consider a simple feedforward network with n input units connected by a weight matrix W to n output linear units. The network is trained to learn the identity function (autoassociation)

from a set of centered training patterns $x_1, \ldots, x_T$. The connection weights are adjusted by gradient descent on the usual LMS error function

$$E(W) = \frac{1}{T} \sum_t \|x_t - Wx_t\|^2 \tag{2.1}$$

The gradient of E with respect to the weights W is given by

$$\nabla E = (W - I)C \tag{2.2}$$

where $C = C_{XX}$ is the covariance matrix of the training set. Thus, the gradient descent learning rule can be expressed as

$$W^{k+1} = W^k - \eta(W^k - I)C \tag{2.3}$$

where W^k is the weight matrix after the kth iteration of the algorithm and η is the constant learning rate ($\eta > 0$). If e_i and $\lambda_i (\lambda_1 \geq \ldots \lambda_n > 0)$ denote the eigenvectors and eigenvalues of C, then

$$W^{k+1}e_i = \eta\lambda_i e_i + (1 - \eta\lambda_i)W^k e_i \tag{2.4}$$

A simple induction shows that

$$W^k = W^0(I - \eta C)^k - [(I - \eta C)^k - I] \tag{2.5}$$

and therefore

$$W^k e_i = [1 - (1 - \eta\lambda_i)^k]e_i + (1 - \eta\lambda_i)^k W^0 e_i \tag{2.6}$$

The behavior of equation 2.6 is clear: provided the learning rate is less than twice the inverse of the largest eigenvalue ($\eta < 2/\lambda_1$), then W^k approaches the identity exponentially fast. This holds for any starting matrix W^0. The eigenvectors of C tend to become eigenvectors of W^k and the corresponding eigenvalues approach 1 at different rates depending on λ_i (larger eigenvalues are learned much faster). As a result, it is not very restrictive to assume, for ease of exposition, that the starting matrix W^0 is diagonal in the e_i basis, i.e., $W^0 = \text{diag}(\alpha_i^{(0)})$ [1] (in addition, learning is often started with the zero matrix). In this case, equation 2.5 becomes

$$W^k e_i = [1 - (1 - \eta\lambda_i)^k(1 - \alpha_i^{(0)})]e_i = \alpha_i^{(k)} e_i \tag{2.7}$$

A simple calculation shows that the corresponding error can be written as

$$E(W^k) = \sum_{i=1}^{n} \lambda_i(\alpha_i^{(k)} - 1)^2 = \sum_{i=1}^{n} \lambda_i(1 - \alpha_i^{(0)})^2(1 - \eta\lambda_i)^{2k} \tag{2.8}$$

[1] Superscripts on the sequence α are in parenthesis to avoid possible confusion with exponentiation.

2.2 Noisy Data. We now modify the setting to introduce noise effects. To fix the ideas, the reader may think for instance that we are dealing with hand-written realizations of single digits numbers. In this case, there are 10 possible patterns but numerous possible noisy realizations. In general, we assume that there is a population of patterns of the form $x_p + n_p$, where x_p denotes the signal and n_p denotes the noise, characterized by the covariance matrices $\bar{C}_{XX}$, $\bar{C}_{NN}$, and $\bar{C}_{XN}$. Here, as everywhere else, we assume that the signal and the noise are centered. A sample $x_t + n_t (1 \leq t \leq T)$ from this population is used as a training set. The training sample is characterized by the covariance matrices $C = C_{XX}$, C_{NN} and C_{XN} calculated over the sample. Similarly, a different sample $x_v + n_v$ from the population is used as a validation set. The validation sample is characterized by the covariance matrices $C' = C'_{XX}$, C'_{NN}, and C'_{XN}. To make the calculations tractable, we shall make, when necessary, several assumptions. First, $C = C = C'$, thus there is a common basis of eigenvectors e_i and corresponding eigenvalues λ_i for the signal in the population and in the training and validation sample. Then, with respect to this basis of eigenvectors, the noise covariance matrices are diagonal $C_{NN} = \text{diag}(\nu_i)$ and $C'_{NN} = \text{diag}(\nu'_i)$. Finally, the signal and the noise are always uncorrelated $C_{XN} = C'_{XN} = 0$. Obviously, it also makes sense to assume that $\bar{C}_{NN} = \text{diag}(\bar{\nu}_i)$ and $\bar{C}_{XN} = 0$ although these assumptions are not needed in the main calculation. Thus we make the simplifying assumptions that both on the training and validation patterns the covariance matrix of the signal is identical to the covariance of the signal over the entire population, the components of the noise are uncorrelated, and the signal and the noise are uncorrelated. Yet we allow the estimates ν_i and ν'_i of the variance of the components of the noise to be different in the training and validation sets.

For a given W, the LMS error function over the training patterns is now

$$E(W) = \frac{1}{T} \sum_t \|x_t - W(x_t + n_t)\|^2 \tag{2.9}$$

By differentiating

$$\nabla E = W(C + C_{NX} + C_{XN} + C_{NN}) - C - C_{XN} \tag{2.10}$$

and since $C_{XN} = C_{NX} = 0$, the gradient is given by

$$\nabla E = (W - I)C + WC_{NN} \tag{2.11}$$

To compute the image of any eigenvector e_i during training, we have

$$W^{k+1} e_i = \eta \lambda_i e_i + (1 - \eta\lambda_i - \eta\nu_i) W^k e_i \tag{2.12}$$

Thus by induction

$$W^k = W^0 M^k - C(C + C_{NN})^{-1}(M^k - I) \tag{2.13}$$

where $M = I - \eta(C + C_{NN})$, and

$$W^k e_i = \frac{\lambda_i}{\lambda_i + \nu_i}[1 - (1 - \eta\lambda_i - \eta\nu_i)^k]e_i + (1 - \eta\lambda_i - \eta\nu_i)^k W^0 e_i \qquad (2.14)$$

Again if we assume here, as in the rest of the paper, that the learning rate satisfies $\eta < \min[1/(\lambda_i + \nu_i)]$, then the eigenvectors of C tend to become eigenvectors of W^k and W^k approaches exponentially fast the diagonal matrix $\mathrm{diag}[\lambda_i/(\lambda_i + \nu_i)]$.[2] Assuming that $W^0 = \mathrm{diag}(\alpha_i^{(0)})$ in the e_i basis, we get

$$W^k e_i = \frac{\lambda_i}{\lambda_i + \nu_i}(1 - b_i a_i^k)e_i = \alpha_i^{(k)} e_i \qquad (2.15)$$

where $b_i = 1 - \alpha_i^{(0)}(\lambda_i + \nu_i)/\lambda_i$ and $a_i = (1 - \eta\lambda_i - \eta\nu_i)$. Notice that $0 < a_i < 1$. Since the signal and the noise are uncorrelated, the error in general can be written in the form

$$E(W) = \frac{1}{P}\sum_p [x_p' x_p - x_p' W x_p - x_p' W' x_p + x_p' W' W x_p + n_p' W' W n_p] \qquad (2.16)$$

Using the fact that $C_{NN} = \mathrm{diag}(\nu_i)$ and $W^k = \mathrm{diag}(\alpha_i^{(k)})$, we have

$$E(W^k) = \sum_{i=1}^{n} [\lambda_i - 2\lambda_i \alpha_i^{(k)} + \lambda_i (\alpha_i^{(k)})^2 + \nu_i (\alpha_i^{(k)})^2] \qquad (2.17)$$

and finally

$$E(W^k) = \sum_{i=1}^{n} [\lambda_i (1 - \alpha_i^{(k)})^2 + \nu_i (\alpha_i^{(k)})^2] \qquad (2.18)$$

It is easy to see that $E(W^k)$ is a monotonically decreasing function of k that approaches an asymptotic residual error value given by

$$E(W^\infty) = \sum_{i=1}^{n} \frac{\lambda_i \nu_i}{(\lambda_i + \nu_i)} \qquad (2.19)$$

For any matrix W, we can define the validation error to be

$$E^V(W) = \frac{1}{V}\sum_v \|x_v - W(x_v + n_v)\|^2 \qquad (2.20)$$

Using the fact that $C'_{XN} = 0$ and $C'_{NN} = \mathrm{diag}(\nu_i')$, a derivation similar to equation 2.18 shows that the validation error $E^V(W^k)$ is given by

$$E^V(W^k) = \sum_{i=1}^{n} [\lambda_i (1 - \alpha_i^{(k)})^2 + \nu_i' (\alpha_i^{(k)})^2] \qquad (2.21)$$

[2] As in equation 2.6, the convergence in fact holds for $\eta < 2\min[1/(\lambda_i + \nu_i)]$. The slightly more restrictive assumption has been chosen to ensure that the numbers a_i are positive.

Clearly, as $k \to \infty$, $E^V(W^k)$ approaches its horizontal asymptote, which is independent of $\alpha_i^{(0)}$ and given by

$$E^V(W^\infty) = \sum_{i=1}^{n} \frac{\lambda_i(\nu_i^2 + \nu_i'\lambda_i)}{(\lambda_i + \nu_i)^2} \tag{2.22}$$

However, it is the behavior of E^V before it reaches its asymptotic value, which is of most interest to us. This behavior, as we shall see, can be fairly complicated.

3 Validation Analysis

Obviously,[3] from equation 2.15, $d\alpha_i^{(k)}/dk = -(\lambda_i b_i a_i^k \log a_i)/(\lambda_i + \nu_i)$. Thus using equation 2.21 and collecting terms yields

$$\frac{dE^V(W^k)}{dk} = \sum_{i=1}^{n} \frac{2\lambda_i^2 b_i \log a_i}{(\lambda_i + \nu_i)^2} a_i^k[\nu_i - \nu_i' + b_i a_i^k(\lambda_i + \nu_i')] \tag{3.1}$$

or, in more compact form,

$$\frac{dE^V}{dk} = \sum_{i=1}^{n} A_i a_i^k + B_i a_i^{2k} \tag{3.2}$$

with

$$A_i = \frac{2\lambda_i^2 b_i}{(\lambda_i + \nu_i)^2}(\nu_i - \nu_i') \log a_i \tag{3.3}$$

and

$$B_i = \frac{2\lambda_i^2 b_i^2}{(\lambda_i + \nu_i)^2}(\lambda_i + \nu_i') \log a_i \tag{3.4}$$

The behavior of E^V depends on the relative size of ν_i and ν_i' and the initial conditions $\alpha_i^{(0)}$, which together determine the signs of b_i, A_i, and B_i. The main result we can prove is as follows.

Assume that learning is started with the zero matrix or with a matrix having sufficiently small weights satisfying, for every i,

$$\alpha_i^{(0)} \leq \min\left(\frac{\lambda_i}{\lambda_i + \nu_i}, \frac{\lambda_i}{\lambda_i + \nu_i'}\right) \tag{3.5}$$

[3]Here and in what follows we take time derivatives with respect to k. Although k was originally introduced as an integer, we can easily consider that $\alpha_i^{(k)}$ and $E^V(W^k)$ are continuous functions of k, defined by equations 2.15 and 2.21, and study them everywhere.

1. If for every i, $\nu_i' \leq \nu_i$, then the validation function E^V decreases monotonically to its asymptotic value and training should be continued as long as possible.

2. If for every i, $\nu_i' > \nu_i$, then the validation function E^V decreases monotonically to a unique minimum and then increases monotonically to its asymptotic value. The derivatives of all orders of E^V have also a unique zero crossing and a unique extremum. For optimal generalization, E^V should be monitored and training stopped as soon as E^V begins to increase. A simple bound on the optimal training time k^{opt} is given by

$$\min_i \frac{1}{\log a_i} \log \frac{-A_i}{B_i} \leq k^{\text{opt}} \leq \max_i \frac{1}{\log a_i} \log \frac{-A_i}{B_i} \tag{3.6}$$

In the most general case of arbitrary initial conditions and noise, the validation function E^V can have several local minima of variable depth before converging to its asymptotic value. The number of local minima is always at most n.

The main result is a consequence of the following statements, which are proved in the Appendix.

First case: For every i, $\nu_i' \geq \nu_i$, i.e., the validation noise is bigger than the training noise. Then

a. If for every i, $\alpha_i^{(0)} \geq \lambda_i/(\lambda_i + \nu_i)$, then E^V decreases monotonically to its asymptotic value.

b. If for every i, $\lambda_i/(\lambda_i + \nu_i') \leq \alpha_i^{(0)} \leq \lambda_i/(\lambda_i + \nu_i)$, then E^V increases monotonically to its asymptotic value.

c. If for every i, $\alpha_i^{(0)} \leq \lambda_i/(\lambda_i + \nu_i')$ and $\nu_i \neq \nu_i'$, then E^V decreases monotonically to a unique global minimum and then increases monotonically to its asymptotic value. The derivatives of all orders of E^V have a unique zero crossing and a unique extremum.

Second case: For every i, $\nu_i' \leq \nu_i$, i.e., the validation noise is smaller than the training noise. Then

a. If for every i, $\alpha_i^{(0)} \geq \lambda_i/(\lambda_i + \nu_i')$ and $\nu_i \neq \nu_i'$, then E^V decreases monotonically to a unique global minimum and then increases monotonically to its asymptotic value. The derivatives of all orders of E^V have a unique zero crossing and a unique extremum.

b. If for every i, $\lambda_i/(\lambda_i + \nu_i) \leq \alpha_i^{(0)} \leq \lambda_i/(\lambda_i + \nu_i')$, then E^V increases monotonically to its asymptotic value.

c. If for every i, $\alpha_i^{(0)} \leq \lambda_i/(\lambda_i + \nu_i)$, then E^V decreases monotonically to its asymptotic value.

Several remarks can be made on the previous statements. First, notice that in both (b) cases, E^V increases because the initial W^0 is already too

good for the given noise levels. The monotone properties of the validation function are not always strict in the sense that, for instance, at the common boundary of some of the cases E^V can be flat. These degenerate cases can be easily checked directly. The statement of the main result assumes that the initial matrix be the zero matrix or a matrix with a diagonal form in the basis of the eigenvectors e_i. A random initial nonzero matrix will not satisfy these conditions. However, E^V is continuous and even infinitely differentiable in all of its parameters. Therefore the results are true also for random sufficiently small matrices. If we use, for instance, an L^2 norm for the matrices, then the norm of a starting matrix is the same in the original or in the orthonormal e_i basis. Equation 3.5 yields a trivial upperbound of $n^{1/2}$ for the norm of the initial diagonal matrix, which roughly corresponds to having random initial weights of order at most $n^{-1/2}$ in the original basis. Thus, heuristically, the variance of the initial random weights should be a decreasing function of the size of the network. This condition is *not* satisfied in many of the usual simulations found in the literature where initial weights are generated randomly and independently using, for instance, a centered gaussian distribution with *fixed* standard deviation. In nonlinear networks, small initial weights are also important for not getting stuck in high local minima during training.

When more arbitrary conditions are considered, in the initial weights or in the noise, multiple local minima can appear in the validation function. As can be seen in one of the curves of the example given in Figure 1, there exist even cases where the first minimum is *not* the deepest one, although these may be rare in some sense, which is not completely understood at this time. In addition, in this particular case, an indication that training should not be stopped at the first minimum comes from the fact that at that point the LMS curve is still decreasing significantly. Also in this figure, better validation results seem to be obtained with smaller initial conditions. This can easily be understood, in this small dimensional example, from some of the arguments given in the Appendix.

Another potentially interesting and relevant phenomena is illustrated in Figure 2. It is possible to have a situation where after a certain number of training cycles, both the LMS and the validation functions appear to be flat and to have converged to their asymptotic values. However, if training is continued, one observes that these plateaux can end and the validation function comes back to life starting to decrease again. In the example, the first minimum is still optimal. However, it is possible to construct examples of validation functions, in higher dimensions, where long plateaux are followed by a phase of significant improvements (see Chauvin 1991).

Finally, we have made an implicit distinction between validation and generalization throughout most of the previous sections. If generalization performance is measured by the LMS error calculated over the entire population, it is clear that our main result can be applied to the generalization error by assuming that $\bar{C}_{NN} = \text{diag}(\bar{\nu}_i)$, and $\nu_i' = \bar{\nu}_i$ for every

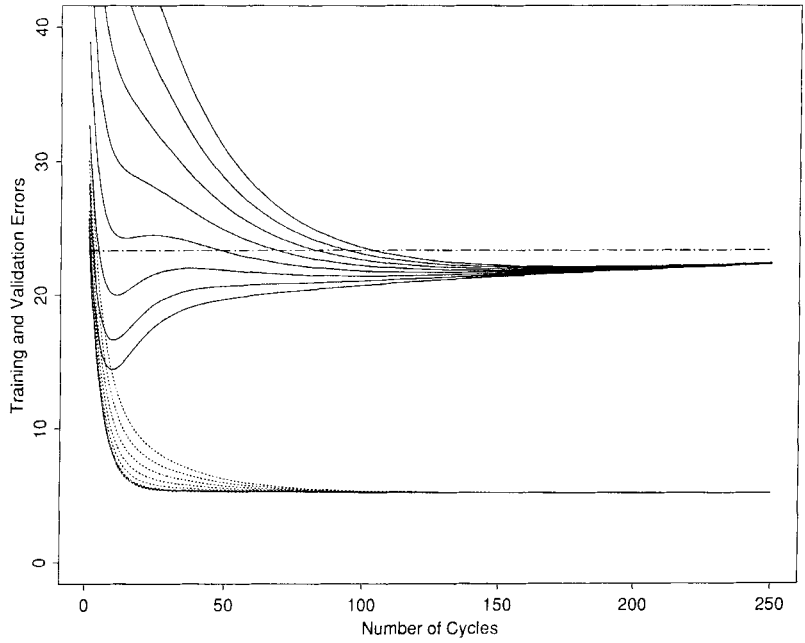

Figure 1: LMS error functions (lower curves) and corresponding validation error functions (upper curves). The parameters are $n = 3$, $\lambda_i = 22, 0.7, 2.5$, $\nu_i = 4, 1, 3$, $\nu_i' = 20, 20, 20$, $\alpha_1^{(0)} = \alpha_2^{(0)} = 0$. From top to bottom, the third initial weight corresponding to $\alpha_3^{(0)}$ takes the values 0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5. The horizontal asymptote of the validation curves is at 23.34. Notice, in particular, the fourth validation curve ($\alpha_3^{(0)} = 0.9$), which has two local minima, the second one being deeper than the first one. At the first minimum, the LMS function is still far from its horizontal asymptote. Also in this case, the validation improves as the initial conditions become closer to 0.

i. In particular, in the second statement of the main result, if for every i $\bar{\nu}_i > \nu_i$, then the generalization curve has a unique minimum. Now, if a validation sample is used as a predictor of generalization performance and the ν_i's are close to the $\bar{\nu}_i$'s, then by continuity the validation and the generalization curves are close to each other. Thus, in this case, the strategy of stopping in a neighborhood of the minimum of the validation function should also lead to near optimal generalization performance.

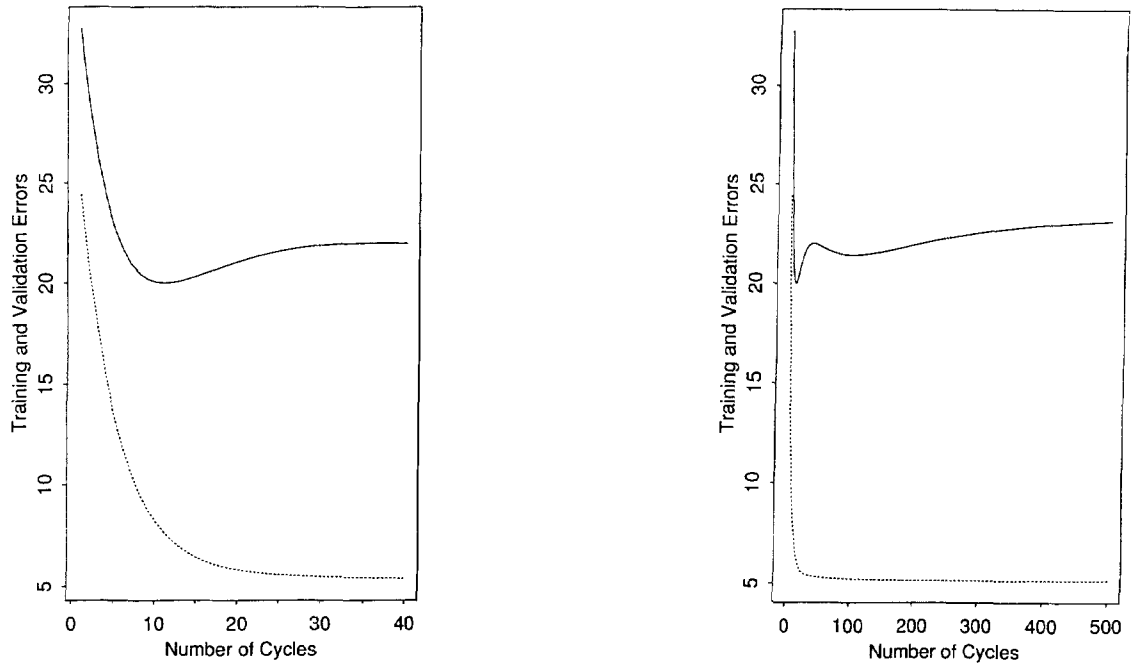

Figure 2: LMS error function (lower curves) and corresponding validation error functions (upper curves). The parameters are $n = 3$, $\lambda_i = 22, 0.7, 2.5$, $\nu_i = 4, 1, 4$, $\nu_i' = 20, 20, 20$, $\alpha_1^{(0)} = \alpha_2^{(0)} = 0$ and $\alpha_3^{(0)} = 0.7$. Notice, on the first two curves, that after 40 cycles both the LMS and the validation function appear to be flat and would suggest one stop the training. The second set of curves corresponds to 500 training cycles. Notice the existence of a second (although shallow) minima, undetectable after 40 cycles.

4 Conclusion

In the framework constructed above, based on linear single layer feedforward networks, it has been possible to analytically derive interesting results on generalization. In particular, under simple noise assumptions, we have given a complete description of the validation error E^V as a function of training time. Although the framework is simplistic, we believe it leads to many nontrivial and perhaps mathematically tractable questions related to generalization. This analysis is only a first step in this direction and many questions remain unanswered. More work is required to test the statistical significance of some of the observations (multiple local minima, plateau effects) and their relevance for practical simulations. For instance, it seems to us that in the case of general noise and arbitrary initial conditions, the upper bound on the number of local minima is rather weak in the sense that, at least on the average, there are many fewer. It seems also that in general the first local minima of E^V is also the deepest. Thus, "pathological" cases may be somewhat rare. In the analysis conducted here, we have used uniform assumptions on the noise. In general, we can expect this not to be the case and properties of the noise cannot be fixed a priori. Therefore one needs to develop a theory of E^V over different possible noise and/or sample realizations, that is to find the average curve E^V (one could also consider averages with respect to initial weights). It would also be of interest to study whether some of the assumptions made on the noise in the training and validation sample can be relaxed and how noise effects can be related to the finite size of the samples. Finally, other possible directions of investigation include the extension to multilayer networks and to general input/output associations.

Appendix: Mathematical Proofs

Let us study E^V under uniform conditions. We shall deal only with the case $\nu_i' \geq \nu_i$ for every i (the case $\nu_i' \leq \nu_i$ is similar).

a. If for every i, $\alpha_i^{(0)} \geq \lambda_i/(\lambda_i + \nu_i)$, then $b_i \leq 0$, $A_i \leq 0$, and $B_i \leq 0$. Therefore, $dE^V/dk \leq 0$ and E^V decreases to its asymptotic value.

b. If for every i, $\lambda_i/(\lambda_i + \nu_i') \leq \alpha_i^{(0)} \leq \lambda_i/(\lambda_i + \nu_i)$, then $0 \leq b_i \leq (\nu_i' - \nu_i)/(\lambda_i + \nu_i')$, $A_i \geq 0$, $B_i \leq 0$, and $A_i + B_i \geq 0$. Since a_i^{2k} decays to 0 faster than a_i^k, $dE^V/dk \geq 0$ and E^V increases its asymptotic value.

c. The most interesting case is when, for every i, $\alpha_i^{(0)} \leq \lambda_i/(\lambda_i + \nu_i')$, i.e., when $b_i \geq (\nu_i' - \nu_i)/(\lambda_i + \nu_i')$. Then $A_i \geq 0$, $B_i \leq 0$, and $A_i + B_i \leq 0$ so that dE^V/dk is negative at the beginning and approaches zero from the positive side as $k \to \infty$. Strictly speaking, this is not satisfied if $A_i = 0$. This can occur only if $b_i = 0$ or $\lambda_i = 0$ (but then $B_i = 0$

also) or if $\nu_i = \nu_i'$. For simplicity, let us add the assumption that $\nu_i \neq \nu_i'$. A function which first increases (respectively decreases) and then decreases (respectively increases) with a unique maximum (respectively minimum) is called unimodal. We need to show that E^V is unimodal. For this, we shall use induction on n combined with an analysis of the unimodality properties of the derivatives of any order of E^V. In fact we will prove the stronger result that the derivatives of all orders of E^V are unimodal and have a unique zero crossing.

For $p = 1, 2, \ldots$, define

$$F^p(k) = \frac{d^p E^V}{dk^p} \tag{4.1}$$

Then

$$F^p(k) = \sum_i f_i^p(k) = \sum_i A_i^p a_i^k + B_i^p a_i^{2k} \tag{4.2}$$

with $A_i^1 = A_i$, $B_i^1 = B_i$, $A_i^p = A_i(\log a_i)^{p-1}$ and $B_i^p = B_i(2\log a_i)^{p-1}$. Clearly, for any $p \geq 1$, sign $(A_i^p) = (-1)^{p+1}$, sign $(B_i^p) = (-1)^p$, and sign $(f_i^p)(0) =$ sign $(A_i^p + B_i^p) = (-1)^p$. Therefore sign $[F^p(0)] = (-1)^p$ and, as $k \to \infty$, $F^p(k)$ approaches zero as $\sum_i A_i^p a_i^k$, thus with the sign of A_i^p which is $(-1)^{p+1}$. As a result, all the continuous functions F^p must have at least one zero crossing. If F^p is unimodal, then F^p has a unique zero crossing. If F^{p+1} has a unique zero crossing, then F^p is unimodal. Thus if for some p_0, F^{p_0} has a unique zero crossing, then all the functions F^p $(1 \leq p < p_0)$ are unimodal and have a unique zero crossing. Therefore, E^V has a unique minimum if and only if there exists an index p such that F^p has a unique zero crossing. By using induction on n, we are going to see that for p large enough this is always the case. Before we start the induction, for any continuously differentiable function f defined over $[0, \infty)$, let

$$\text{zero}(f) = \inf\{x : f(x) = 0\} \tag{4.3}$$

and

$$\text{ext}(f) = \inf\left\{x : \frac{df}{dx}(x) = 0\right\} \tag{4.4}$$

Most of the time, zero and ext will be applied to functions that in fact have a unique zero or extremum. In particular, for any i and p, it is trivial to see that the functions f_i^p are unimodal and with a unique zero crossing. A simple calculation gives

$$\text{zero}(f_i^p) = \frac{1}{\log a_i} \log \frac{-A_i}{2^{p-1}B_i} = \frac{1}{\log a_i} \log \frac{\nu_i' - \nu_i}{2^{p-1}b_i(\lambda_i + \nu_i')} \tag{4.5}$$

and

$$\text{ext}(f_i^p) = zero(f_i^{p+1}) = \frac{1}{\log a_i} \log \frac{-A_i}{2^p B_i} = \frac{1}{\log a_i} \log \frac{\nu_i' - \nu_i}{2^p b_i(\lambda_i + \nu_i')} \tag{4.6}$$

Also notice that for any $p \geq 1$

$$\min_i \text{zero}(f_i^p) \leq \text{zero} F^p \leq \max_i \text{zero}(f_i^p) \tag{4.7}$$

and

$$\min_i \text{ext}(f_i^p) \leq \text{ext} F^p \leq \max_i \text{ext}(f_i^p) \tag{4.8}$$

(equations 4.7 and 4.8 are in fact true for *any* zero crossing or extremum of F^p).

We can now begin the induction. For $n = 1$, E^V has trivially a unique minimum and all its derivatives are unimodal with a unique zero crossing. Let us suppose that this is also true of any validation error function of $n - 1$ variables. Let $\lambda_1 \geq \ldots \geq \lambda_n > 0$ and consider the corresponding ordering induced on the variables $a_i = 1 - \eta\lambda_i - \eta\nu_i$, $1 > a_{i_1} \geq \ldots a_{i_n} \geq 0$. Let i_j be a fixed index such that $a_{i_1} \geq a_{i_j} \geq a_{i_n}$ and write, for any $p \geq 1$, $F^p(k) = G^p(k) + f_{i_j}^p(k)$ with $G^p(k) = \sum_{i \neq i_j} f_i^p(k)$. $f_{i_j}^p$ is unimodal with a unique zero crossing and so is G^p by the induction hypothesis. Now it is easy to see that F^p will have a unique zero crossing if

$$\text{zero}(G^p) \leq \text{zero}(f_{i_j}^p) \leq \text{ext}(G^p) \tag{4.9}$$

By applying equations 4.7 and 4.8 to G^p, we see that F^p will have a unique zero crossing if

$$\max_{i \neq i_j} \text{zero}(f_i^p) \leq \text{zero}(f_{i_j}^p) \leq \min_{i \neq i_j} \text{ext}(f_i^p) \tag{4.10}$$

Substituting the values given by equations 4.5 and 4.6, we can see that for large p, equation 4.10 is equivalent to

$$\max_{i \neq i_j} -p\frac{\log 2}{\log a_i} \leq -p\frac{\log 2}{\log a_{i_j j}} \leq \min_{i \neq i_j} -p\frac{\log 2}{\log a_i} \tag{4.11}$$

and this is satisfied since $a_{i_1} \geq \cdots \geq a_{i_n}$. Therefore, using the induction hypothesis, we see that there exists an integer p_0 such that, for any $p > p_0$, F^p has has a unique zero crossing. But, as we have seen, this implies that F^p has a unique zero crossing also for $1 \leq p \leq p_0$. Therefore E^V is unimodal with a unique minimum and its derivatives of all orders are unimodal with a unique zero crossing.

Notice that $F(k)$ cannot be zero if all the functions $f_i(k)$ are simultaneously negative or positive. Therefore, a simple bound on the position of the unique minimum k^{opt} is given by

$$\min_i \text{ zero}(f_i) \leq \text{zero}(F) \leq \max_i \text{ zero}(f_i) \tag{4.12}$$

or

$$\min_i \frac{1}{\log a_i} \log \frac{-A_i}{B_i} \leq k^{\text{opt}} \leq \max_i \frac{1}{\log a_i} \log \frac{-A_i}{B_i} \tag{4.13}$$

[It is also possible, for instance, to study the effect of the initial $\alpha_i^{(0)}$ on the position or the value of the local minima. By differentiating the relation $F^1(k) = 0$ one gets immediately

$$F^2(k)dk = \sum_i \left(\frac{\lambda_i + \nu_i}{\lambda_i b_i} \right) (A_i a_i^k + 2B_i a_i^{2k}) d\alpha_i^{(0)} \tag{4.14}$$

(see Fig. 2)].

To find an upper bound on the number of local minima of E^V in the general case of arbitrary noise and initial conditions, we first order the $2n$ numbers a_i and a_i^2 into an increasing sequence c_i, $i = 1, \ldots, 2n$. This induces a corresponding ordering on the $2n$ numbers A_i and B_i yielding a second sequence C_i, $i = 1, \ldots, 2n$. Now the derivative of E^V can be written in the form

$$\frac{dE^V}{dk} = F^1(k) = \int C(a) a^k d\mu(a) \tag{4.15}$$

where μ is the finite positive measure concentrated at the points a_i and a_i^2. The kernel a^k in the integral is totally positive. Thus (see, for instance, Karlin 1968, theorem 3.1, p. 233) the number of sign changes of $F^1(k)$ is bounded by the number of sign changes in the sequence C. Therefore the number of sign changes in F^1 is at most $2n - 1$ and the number of zeros of F^1 is at most $2n - 1$. So the number of local minima of E^V is at most n.

Acknowledgments

This work is in part supported by grants from the Office of Naval Research and the McDonnell-Pew foundation to P. B. We would like to thank Yosi Rinott for useful discussions.

References

Baldi, P., and Hornik, K. 1989. Neural network and principal component analysis: Learning from examples without local minima. *Neural Networks* **2**, 53–58.

Baldi, P., and Hornik, K. 1991. Back-propagation and unsupervised learning in linear networks. In *Back-propagation: Theory, Architectures and Applications*, Y. Chauvin and D. E. Rumelhart, eds. Lawrence Erlbaum, NJ. In press.

Baum, E. B., and Haussler, D. 1989. What size net gives valid generalization? *Neural Comp.* **1**, 151–160.

Chauvin, Y. 1991. Generalization dynamics in LMS trained linear networks. *Neural Information Processing Systems 3* (Proceedings of the 1990 NIPS Conference). Morgan Kaufmann, San Mateo, CA.

Karlin, S. 1968. *Total Positivity.* Stanford University Press. Stanford, CA.

Sompolinsky, H., Tishby, N., and Seung, H. S. 1990. Learning from examples in large neural networks. *Phys. Rev. Lett.* **65**(13), 1683–1686.

Tishby, N., Levin, E., and Solla, S. A. 1989. Consistent inference of probabilities in layered networks: Predictions and generalization. In *Proceedings of the IJCNN*, pp. 403–409. IEEE, New York.

Received 1 February 1991; accepted 13 April 1991.