Sensitivity of Feedforward Neural Networks to Weight Errors

MARYHELEN STEVENSON, STUDENT MEMBER, IEEE, RODNEY WINTER, MEMBER, IEEE, AND BERNARD WIDROW, FELLOW, IEEE

Abstract—**An important consideration when implementing neural networks with digital or analog hardware of limited precision is the sensitivity of neural networks to weight errors. In this paper, we analyze the sensitivity of feedforward layered networks of Adaline elements (threshold logic units) to weight errors. An approximation is derived which expresses the probability of error for an output neuron of a large network (a network with many neurons per layer) as a function of the percentage change in the weights. As would be expected, the probability of error increases with the number of layers in the network and with the percentage change in the weights. Surprisingly, the probability of error is essentially independent of the number of weights per neuron and of the number of neurons per layer, as long as these numbers are large (on the order of 100 or more).**

I. Introduction

THE input–output function realized by a neural network is determined by the values of its weights. Recently, a great deal of effort has been devoted to developing algorithms which will adapt the weights to realize a desired input–output mapping [1], [2]. When using limited-precision hardware to store the desired weights, an important issue is that of weight sensitivity; how sensitive is the input–output mapping of the neural network to weight errors? We will investigate this question for feedforward networks of Adaline elements.

The outline of the paper is as follows. The network of Adalines, called Madaline [1], [3], [4], is described in Section II. Notation and terminology from n-dimensional geometry is introduced in Section III. Section IV describes the Hoff hypersphere-area approximation which is used for the sensitivity analysis. In Section V, we determine the sensitivity of an Adaline (the basic unit of the Madaline) to weight errors. Since output errors from one layer become input errors to the following layer, it is also necessary to determine the sensitivity of an Adaline to input errors; this is done in Section VI. In Section VII, we establish a method for determining the sensitivity of an Adaline to a combination of weight errors and input errors. Experimental results which support our theory on Adaline sensitivity are presented in Section VIII. Section IX uses the results on Adaline sensitivity to determine the sensitivity of a Madaline to weight errors. Finally, in Section X, we present experimental results which support the theory developed in Section IX. In this manner, the sensitivity of a layered neural network to weight errors is determined.

Manuscript received August 23, 1989; revised October 31, 1989. This work was supported by SDIO Innovative Science and Technology Office and managed by ONR under Contract N00014-86-K-0718, and by the Department of the Army Belvoir RD&E Center under Contract DAAK 70-89-K-0001, by NASA under Contract NCA2-389, by Rome Air Development Center under Contract F30602-88-D-0025, Subcontract E-21-T22-S1, and by a grant from the Lockheed Missiles and Space Company.

M. Stevenson and B. Widrow are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305.

R. Winter is with the U.S. Air Force PSC, APO NY 09194-5421.

IEEE Log Number 8933163.

II. The Madaline Architecture

The *Adaline* (*ada*ptive *lin*ear *e*lement) [3], [5] (also known as a linear threshold unit) is the basic building block of the *Madaline* (*m*any *Adaline*s) network. Fig. 1 shows an Adaline with n variable inputs: $x_1, x_2, \cdots, x_n$. The inputs take on binary values of either $+1$ or -1. The bias input x_0 is fixed at a value of $+1$. Associated with the Adaline are $n+1$ adjustable analog weights: $w_0, w_1, \cdots, w_n$. The weights of the Adaline scale the corresponding inputs, the scaled inputs are summed, and the weighted sum is the input to a threshold device. The threshold device outputs a -1 for negative inputs and a $+1$ for positive inputs. The output of the threshold device is the Adaline output. The input–output map of the Adaline can be summarized as:

$$\text{output of Adaline} = \begin{cases} 1, & \sum_{i=0}^{n} x_i w_i \geq 0 \\ -1, & \sum_{i=0}^{n} x_i w_i < 0. \end{cases} \quad (1)$$

A layered network of Adaline elements (a Madaline) is shown in Fig. 2. The inputs to the network are presented to each of the Adalines in the first layer. The outputs from the first-layer Adalines then serve as inputs to the second-layer Adalines, and so on. The Adalines of the final layer (in this case, the third layer) are called the output Adalines. Their outputs are the outputs of the network.

III. Review of n-Dimensional Geometry

The use of n-dimensional geometry has proven to be a valuable tool for understanding and analyzing the Adaline. The geometrical interpretation of the equation dictating the Adaline's input–output map is the basis for most of the analysis presented in this paper. A good reference

1045-9227/90/0300-0071$01.00 © 1990 IEEE

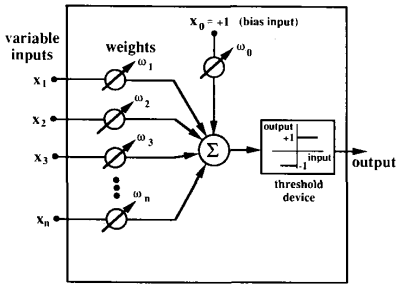


Fig. 1. The adaptive linear element, or Adaline.

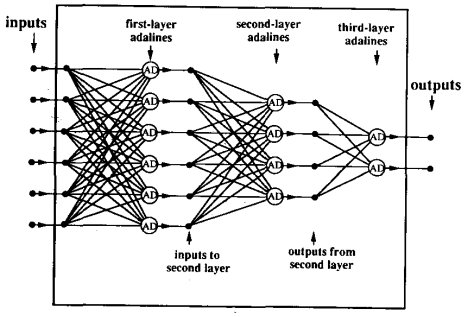


Fig. 2. A three-layer Madaline.

on the geometry of n dimensions is [6]. The essential notation and terminology that we use in this paper is presented below.

- The *vector* from the origin to the point $(x_0, x_1, \cdots, x_n)$ in $(n+1)$-space is denoted by X. We will refer to the point $(x_0, x_1, \cdots, x_n)$ as "the tip of X."
- The collection of all points in n space which are at a distance r from the point c is a *hypersphere* of radius r centered at c. The surface area[1] of a hypersphere of radius r in n space $A_n(r)$ is:

$$A_n(r) = K_n r^{n-1} \tag{2}$$

where

$$K_n = 2\pi^{n/2}/\Gamma(n/2) \tag{3}$$

and $\Gamma(\cdot)$ is the well-known Gamma function.

- Given a vector $W \triangleq [w_0, w_1, \cdots, w_n]$, the collection of all points $(x_0, x_1, \cdots, x_n)$ in $(n+1)$ space which satisfy

$$X \cdot W \triangleq \sum_{i=0}^{n} x_i w_i = c \tag{4}$$

for some scalar c, is called a *hyperplane*. This hyperplane is perpendicular to the vector W and is at a distance $c/|W|$ from the origin (where $|W|$ is the magnitude of W). We use the notation HP_W to denote the hyperplane perpendicular to W which passes through the origin. Note that HP_W is the hyperplane described by (4) when $c = 0$:

$$\mathrm{HP}_W \triangleq \left\{ (x_0, x_1, \cdots, x_n) : X \cdot W = \sum_{i=0}^{n} x_i w_i = 0 \right\}. \tag{5}$$

- Any hyperplane which passes through the center of a hypersphere divides the hypersphere in two *hemihyperspheres*. Let X be a vector in $(n+1)$ space and let HP_X denote the hyperplane which passes through the origin and which is perpendicular to X. Then HP_X divides any hypersphere centered at the origin in two hemihyperspheres which we call H_X^+ and H_X^-. We use the notation H_X^+ to denote the hemihypersphere on the $+X$ side of HP_X and H_X^- to denote the hemihypersphere on the $-X$ side of HP_X.
- A *lune* is the section of a hypersphere sandwiched between designated sides of two hyperplanes both of which pass through the center of the hypersphere (see Fig. 3). Let X and W be $(n+1)$-dimensional vectors. Consider any hypersphere centered at the origin in $(n+1)$ space. The hyperplane HP_X divides the hypersphere into the hemihyperspheres H_X^+ and H_X^- while the hyperplane HP_W divides the hypersphere into the hemihyperspheres H_W^+ and H_W^-. If the angle between X and W is θ, then the intersections $H_X^+ \cap H_W^-$ and $H_X^- \cap H_W^+$ both describe lunes of angle θ whereas the intersections $H_X^+ \cap H_W^+$ and $H_X^- \cap H_W^-$ both describe lunes of angle $(\pi - \theta)$. The ratio of the surface content of a lune of angle θ to the surface content of the entire hypersphere is $\theta/(2\pi)$.

The geometric interpretation of the equation dictating the input–output map of an Adaline (1) provides insight concerning the manner in which an Adaline's weight vector dichotomizes the input space. For a given weight vector W

$$X \cdot W = \sum_{i=0}^{n} x_i w_i = 0 \tag{6}$$

is the equation of the hyperplane HP_W. This hyperplane, sometimes referred to as the *separating hyperplane*, separates the X for which $X \cdot W > 0$ from those X for which $X \cdot W < 0$ (i.e., it separates those input vectors which yield a $+1$ response from those input vectors which yield a -1 response). Geometrically, we see that the output of the Adaline is determined by the angle between the input vector and the weight vector[2]; the output is $+1$ if the angle between the two vectors is less than 90° and is -1 if this angle is greater than 90°.

IV. The Hoff Hypersphere-Area Approximation

Assuming binary-valued (± 1) inputs, there are 2^n possible input patterns for an Adaline with n variable inputs. Each input pattern corresponds to a point in n space which

[1]To be technically correct we should say surface content rather than surface area.

[2]Note that if we had not included the bias input and bias weight as components of the input vector and weight vector, respectively, but had instead used the n-dimensional vectors with $-w_0$ replacing 0 as the threshold for the threshold device, the separating hyperplane would no longer pass through the origin and the output of the Adaline would no longer depend solely on the angle between the n-dimensional input and weight vectors.

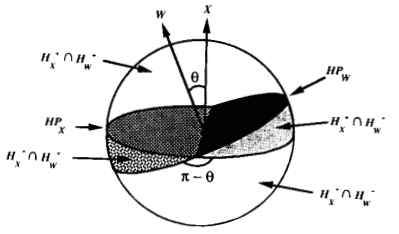


Fig. 3. The hyperplane HP_X divides the hypersphere in two hemihyperspheres H_X^+ and H_X^-. Similarly, the hyperplane HP_W divides the hypersphere in two hemihyperspheres H_W^+ and H_W^-. The section of the hypersphere on the $+X$ side of HP_X and on the $-W$ side of HP_W is a lune of angle θ designated as $H_X^+ \cap H_W^-$.

lies on a hypersphere of radius $n^{1/2}$ centered at the origin. The Hoff hypersphere-area approximation states that as n gets large, the points corresponding to the n-dimensional input patterns are approximately uniformly distributed over the surface of a hypersphere in n space. Consequently, the percentage of input patterns which correspond to points on a selected region of the hypersphere can be approximated as the ratio of the surface content of the selected region to the surface content of the entire hypersphere. The validity of this approximation was shown by Hoff in his doctoral dissertation [7]. As will be seen, the hypersphere-area approximation is an extremely useful tool for analyzing the expected behavior of the Adaline.

The hypersphere-area approximation requires a slight modification if we include the bias input as a component of the input vector. The reason for this is that the bias input, x_0, is always $+1$; there are no input vectors with $x_0 = -1$. This means that the points corresponding to the $(n + 1)$-dimensional input vectors are distributed over only half of a hypersphere (the half corresponding to $x_0 > 0$). Glanz [8] modified the hypersphere-area approximation for use in $(n + 1)$ space. The hypersphere-area approximation as modified by Glanz states that for large n, the points corresponding to the $(n + 1)$-dimensional input vectors are approximately uniformly distributed over a hemihypersphere of radius $(n + 1)^{1/2}$ in $(n + 1)$ space.

V. Adaline Errors Due to Weight Perturbations

The weight vector determines the input–output map of an Adaline. A slight change in the direction of the weight vector can alter this map. In this section, we study the effect of a weight vector perturbation on the input–output map of an Adaline. More specifically, we find the probability that two weight vectors (one considered as a perturbation of the other) map an arbitrary input vector into opposite output categories. This is the probability of an Adaline decision error due to the given weight perturbation. The main result of this section is an expression for the probability of an Adaline decision error as a function of the weight perturbation ratio.

Consider an Adaline and its associated weight vector W. If a randomly oriented *perturbation vector* ΔW is added to W, then the resulting vector $W_p = W + \Delta W$ is the *perturbed weight vector*. The *weight perturbation ratio*, denoted by δW, is defined to be the ratio of the magnitude of the perturbation vector to the magnitude of the original weight vector:

$$\delta W \triangleq |\Delta W| / |W|. \tag{7}$$

Let θ_{WW_p} denote the *angular perturbation* (i.e., the angle between W and W_p). We will first show that the probability of a decision error is proportional to the angular perturbation and then establish the relation between the weight perturbation ratio and the angular weight perturbation.

Let HP_W be the hyperplane perpendicular to W which passes through the origin. This hyperplane divides the *input hypersphere* (the hypersphere on which the tips of the input vectors lie) into the two hemihyperspheres: H_W^+ and $H_{\bar{W}}$. All input vectors whose tips lie on H_W^+ have a positive dot product with W and all input vectors whose tips lie on $H_{\bar{W}}$ have a negative dot product with W. Thus, an Adaline with weight vector W maps all input vectors corresponding to points on H_W^+ to $+1$ and maps all input vectors corresponding to points on $H_{\bar{W}}$ to -1. Similarly, the hyperplane HP_{W_p} divides the input hypersphere into the two hemihyperspheres: $H_{W_p}^+$ and $H_{\bar{W}_p}$. An Adaline having W_p as its weight vector will map all input vectors with tips on $H_{W_p}^+$ to $+1$ and will map all input vectors with tips on $H_{\bar{W}_p}$ to -1.

The probability that an arbitrary input vector is mapped into opposite output categories by W and W_p is the fraction of input vectors with tips either on the intersection of H_W^+ and $H_{\bar{W}_p}$ or on the intersection of $H_{\bar{W}}$ and $H_{W_p}^+$. Both of these intersections describe lunes of angle θ_{WW_p}. Note that these two lunes are spherical reflections of each other (i.e., the reflection through the origin of a point on the first lune results in a point on the second lune and vice versa). Therefore, if a third hyperplane HP_{X_0} is introduced, which also passes through the origin and which is randomly oriented with respect to HP_W and HP_{W_p}, then exactly half of the combined surface area of the two lunes will lie on the hemihypersphere $H_{X_0}^+$ and half will lie on the hemihypersphere $H_{\bar{X}_0}$. The purpose of considering this third hyperplane is to account for the fact that the input vectors are distributed over only half of the input hypersphere. Applying the hypersphere-area approximation as modified by Glanz, the fraction of input vectors mapped into different categories by W and W_p is computed as the fraction of the surface area of the hemihypersphere $H_{X_0}^+$, which belongs to one of the two lunes described above. This fraction is given by the ratio of the surface area of a lune of angle θ_{WW_p} to the surface area of the hemihypersphere on which it lies. So the probability of a decision error due to an angular weight perturbation of θ_{WW_p} is:

$$P(\text{Decision Error}) = \theta_{WW_p}/\pi. \tag{8}$$

For practical applications, the probability of a decision error should be expressed in terms of the weight perturbation ratio as opposed to the angular weight perturba-

tion. Given a weight perturbation ratio, the angular perturbation is a random variable which, depending on the orientation of the perturbation vector with respect to the original weight vector, will have a value in the range: $0 \leq \theta_{WW_p} \leq \sin^{-1}(\delta W)$. In this case, the probability of a decision error is computed by replacing θ_{WW_p} in (8) by its expected value, $\overline{\theta_{WW_p}}$:

$$P(\text{Decision Error}) = \overline{\theta_{WW_p}}/\pi. \tag{9}$$

The next task is to find an expression for $\overline{\theta_{WW_p}}$ in terms of the weight perturbation ratio. The geometry of the problem is depicted in Fig. 4 where ϕ denotes the unknown angle between W and ΔW. From the figure, we see that θ_{WW_p} can be expressed in terms of ϕ, $|W|$, and $|\Delta W|$ as follows:

$$\theta_{WW_p} = \tan^{-1}\left(|\Delta W| \sin\phi/(|W| + |\Delta W| \cos\phi)\right) \tag{10}$$

$$= \tan^{-1}\left(\delta W \sin\phi/(1 + \delta W \cos\phi)\right) \tag{11}$$

$$\approx \tan^{-1}(\delta W \sin\phi), \quad \text{for } \delta W \ll 1. \tag{12}$$

Using $\tan^{-1} x \approx x$ for small x

$$\theta_{WW_p} \approx \delta W \sin\phi, \quad \text{for small } \delta W. \tag{13}$$

We now have an expression for the random variable θ_{WW_p} in terms of the weight perturbation ratio and the random variable ϕ. Assuming the weight perturbation ratio is known, we find the expected value of θ_{WW_p} to be:

$$\overline{\theta_{WW_p}} \approx \delta W E\{\sin\phi\}, \quad \text{for small } \delta W. \tag{14}$$

Glanz [8] showed that the probability density function for the random variable ϕ (the angle between two randomly oriented variables in $(n+1)$ space) is given by

$$f_\phi(\varphi) = (K_n/K_{n+1}) \sin^{n-1}\varphi, \quad 0 < \varphi < \pi \tag{15}$$

where K_n is defined in (3). Using this probability density function for the random variable ϕ, Winter [9] found that for large n, the expected value of $\sin\phi$ is very close to one. Briefly, his steps were as follows:

$$E\{\sin\phi\} = \int_0^\pi \sin\varphi \underbrace{(K_n/K_{n+1}) \sin^{n-1}\varphi}_{\text{pdf of } \phi} d\varphi$$

$$= (K_n/K_{n+1})^2 2\pi/n. \tag{16}$$

Stirling's approximation can be used to show that:

$$K_n/K_{n+1} \approx (2\pi/n)^{-1/2}, \quad \text{for large } n. \tag{17}$$

So for large n, $E\{\sin\phi\} \approx 1$ and

$$\overline{\theta_{WW_p}} \approx \delta W, \quad \text{for small } \delta W. \tag{18}$$

Substituting this expression for $\overline{\theta_{WW_p}}$ in (9), we conclude that the probability of an Adaline decision error resulting from a given weight perturbation ratio can be approximated as the weight perturbation ratio divided by π:

$$P[\text{Decision Error}] \approx \delta W/\pi. \tag{19}$$

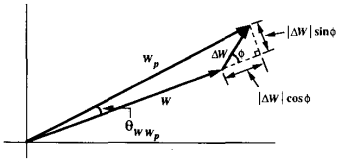


Fig. 4. For a given weight perturbation ratio, the value of θ_{WW_p} depends on the value of the random variable ϕ: $\theta_{WW_p} = \tan^{-1}(|\Delta W| \sin\phi/(|W| + |\Delta W| \cos\phi))$.

This approximation is based on the assumptions that n is large and δW is small.

VI. Decision Errors Due to Input Errors

In a network of Adalines, the outputs from the Adalines of one layer are the inputs to the Adalines of the next layer (see Fig. 2). This means that Adaline decision errors of one layer become input errors to the Adalines of the following layer. For this reason, it is necessary to understand the effect of input errors on the input–output mapping of an Adaline. In this section, we find the probability that two input vectors X and X_p are mapped into opposite output categories by an arbitrary weight vector W. Although X_p is itself an input vector, we view it as a perturbation of the input vector X and say that a decision error occurs whenever X and X_p are mapped into different categories.

Let X and X_p be input vectors of dimension $(n+1)$. The component of X and X_p which corresponds to the bias input is fixed at a value of $+1$; the other n components of X and X_p are variables which take on values of either $+1$ or -1. The *Hamming distance* between two binary-valued vectors is defined to be the number of components for which the two vectors differ in value. If X and X_p are separated by a Hamming distance of h $(0 \leq h \leq n)$, then the vector X_p can be thought of as the vector X with h errors. The probability that X and X_p are mapped into opposite output categories by an arbitrary weight vector W is the probability of an Adaline decision error due to h input errors. If the tip of W is drawn from a uniform distribution over the surface of a hypersphere centered at the origin in $n+1$ space, then it is straightforward to show that this probability is given by θ_{XX_p}/π, where θ_{XX_p} is the angle between the angle between X and X_p:

$$P(\text{Decision Error}) = \theta_{XX_p}/\pi. \tag{20}$$

The angle θ_{XX_p} is easily computed using the fact that the dot product of two vectors is equal to the product of the magnitude of the two vectors and the cosine of the angle between them:

$$\theta_{XX_p} = \cos^{-1}\left((X \cdot X_p)/(|X|\,|X_p|)\right). \tag{21}$$

Both X and X_p have magnitude $(n+1)^{1/2}$ and the dot product of the two vectors is $n + 1 - 2h$, where h is the Hamming distance between X_p and X:

$$\theta_{XX_p} = \cos^{-1}\left(1 - 2h/(n+1)\right). \tag{22}$$

Substituting this expression for θ_{XX_p} in (20), we find the probability that two input vectors, separated by a Hamming distance of h, are mapped into opposite output categories by an arbitrary weight vector:

$$P(\text{Decision Error}) = (1/\pi)\cos^{-1}\left(1 - 2h/(n+1)\right) \quad (23)$$

$$\approx (1/\pi)\left[4h/(n+1)\right]^{1/2} \quad (h \ll n+1) \quad (24)$$

where the last approximation uses $\cos(\theta) \approx 1 - \theta^2/2$, for small θ.

For the purpose of expressing the probability of a decision error in terms of the input perturbation ratio, imagine that the inversion of h components of X is accomplished by adding a perturbation vector ΔX to X:

$$X_p = X + \Delta X. \quad (25)$$

Let x_i, x_{p_i}, and Δx_i denote the ith component of X, X_p, and ΔX, respectively. Then

$$\Delta x_i = \begin{cases} 0, & x_{p_i} = x_i \\ -2x_i, & x_{p_i} = -x_i. \end{cases} \quad (26)$$

Since X and X_p differ in h components, the magnitude of ΔX is $(4h)^{1/2}$. So the *input perturbation ratio* δX resulting from h input errors is:

$$\delta X \triangleq |\Delta X|/|X| = \left[4h/(n+1)\right]^{1/2}. \quad (27)$$

Comparing this expression with the approximation for the probability of a decision error due to h input errors (approximation (24)), we see that the probability of an Adaline decision error due to an input perturbation ratio of δX is approximately:

$$P(\text{Decision Error}) \approx \delta X/\pi. \quad (28)$$

We conclude (compare (28) and (19)) that the probability of a decision error is approximated as the perturbation ratio divided by π; it makes no difference whether the perturbation ratio describes an input perturbation or a weight perturbation.

VII. Decision Errors in the Presence of Both Weight Perturbations and Input Errors

In the previous two sections, we have found the probability that an Adaline makes an error due to either a weight perturbation or to input errors. However, if both the weight and input vectors are perturbed then it is not obvious how to determine the net effect of the two perturbations. In this section we establish a method for finding the probability of an Adaline decision error due to a combination of weight perturbations and input errors.

Comparing the results of the last two sections, it is clear that a given perturbation ratio results in the same error probability regardless of whether it refers to an input perturbation or a weight perturbation. This suggests that when both types of perturbation are present, the probability that the perturbed Adaline makes a decision error (with respect to the original Adaline) can be found by considering both perturbations to be of the same type and then finding the expected net perturbation ratio.

Consider an Adaline with both input and weight perturbations. We can describe each of the perturbations in terms of the angle between the perturbed and unperturbed vectors. Let θ_{XX_p} be the angle between X and X_p, and let θ_{WW_p} be the angle between W and W_p. Since both perturbation types have the same effect on the probability of error, the input perturbation can be considered as an additional weight perturbation (see Fig. 5). To do this, we first perturb (rotate) the original weight vector by the angle θ_{WW_p} and call the resulting vector W_p. Next, we perturb W_p by the angle θ_{XX_p}. This results in the doubly perturbed weight vector W_{pp}. The net angular perturbation to the weight vector is given by $\theta_{WW_{pp}}$, the angle between W and W_{pp}. Depending on the relative directions of the perturbations, the net angular perturbation will vary between $|\theta_{WW_p} - \theta_{XX_p}|$ and $|\theta_{WW_p} + \theta_{XX_p}|$. Let $\overline{\theta_{WW_{pp}}}$ denote the expected value of the net angular perturbation given both the weight perturbation ratio and the input perturbation ratio. Then the probability that the perturbed Adaline makes a decision error with respect to the original Adaline (i.e., the probability that the output category to which W maps X is opposite to the output category to which W_p maps X_p) is:

$$P(\text{Decision Error}) = \overline{\theta_{WW_{pp}}}/\pi. \quad (29)$$

An expression for $\overline{\theta_{WW_{pp}}}$ must now be found in terms of the input and weight perturbation ratios. For small perturbation ratios, the angular perturbation is approximately equal to the perturbation ratio. Thus, we will find the expected value of the net perturbation ratio and use this as the expected value of the net angular perturbation. For this purpose, consider two randomly oriented vectors P_1 and P_2, of fixed magnitudes in $(n+1)$ space. Let ϕ be a random variable which denotes the angle between P_1 and P_2, and let P be the vector sum of P_1 and P_2. Then the magnitude of P is a random variable and can be expressed as a function of the random variable ϕ as shown below:

$$|P| = \left(|P_1|^2 + |P_2|^2 - 2|P_1|\,|P_2|\cos(\pi - \phi)\right)^{1/2}. \quad (30)$$

The expected value of $|P|$ is found by multiplying the expression for the magnitude of P by the probability density function of ϕ (see (15)) and integrating over the range of ϕ:

$$E(|P|) = \int_0^{\pi} \underbrace{(K_n/K_{n+1})\sin^{n-1}(\varphi)}_{\text{pdf of } \phi} \left(|P_1|^2 + |P_2|^2 - 2|P_1|\,|P_2|\cos(\pi - \varphi)\right)^{1/2} d\varphi. \quad (31)$$

For large n, $\sin^{n-1}(\varphi)$ is close to 0, except when φ is close to $\pi/2$. This means that for large values of n, there is a high probability that ϕ has a value close to $\pi/2$. When

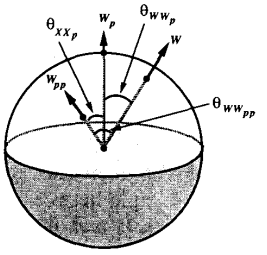


Fig. 5. The input perturbation can be considered as an additional weight perturbation. The original weight vector W is first perturbed (rotated) by the weight perturbation angle θ_{WW_p}, and then perturbed by the input perturbation angle θ_{XX_p}. This results in the doubly perturbed weight vector W_{pp}. The net angular perturbation is given by $\theta_{WW_{pp}}$, the angle between W and W_{pp}.

$\varphi = \pi/2$, $\cos(\pi - \varphi) = 0$. Hence, the integral above can be approximated as:

$$E(|P|) \approx \left(|P_1|^2 + |P_2|^2\right)^{1/2} \cdot \underbrace{\int_0^{\pi} (K_n/K_{n+1}) \sin^{n-1}(\varphi)\, d\varphi}_{1} \quad (32)$$

$$= \left(|P_1|^2 + |P_2|^2\right)^{1/2}. \quad (33)$$

Equation (33) can be used to determine the expected magnitude of the net perturbation resulting from two independent perturbations. Let δX be the input perturbation ratio and let δW be the weight perturbation ratio. The plan is to regard both perturbations as weight perturbations. The input perturbation is converted to an equivalent weight perturbation by scaling the input perturbation ratio by the magnitude of the weight vector. Using (33), the expected magnitude of the net perturbation is:

$$E(\text{magnitude of net perturbation}) \approx \left[(\delta X|W|)^2 + (\delta W|W|)^2\right]^{1/2}. \quad (34)$$

So the expected net perturbation ratio is approximated as the square root of the sum of the squares of the input perturbation ratio and the weight perturbation ratio:

$$E(\text{net perturbation ratio}) \approx \left[(\delta X)^2 + (\delta W)^2\right]^{1/2}. \quad (35)$$

Using this approximation for the expected net perturbation ratio as an approximation for the expected net angular perturbation and substituting into (29), we find that the probability of an Adaline decision error due to an input perturbation of δX and a weight perturbation ratio of δW is:

$$P(\text{Decision Error}) \approx (1/\pi)\left[(\delta X)^2 + (\delta W)^2\right]^{1/2}. \quad (36)$$

VIII. Simulation Results on Adaline Sensitivity

A computer simulation was run to determine the relative frequency of an Adaline error as a function of various combinations of weight and input perturbations. To do so, a reference weight vector and a reference input vector were randomly chosen. The weight vector was chosen from a uniform distribution over the surface of a hypersphere centered at the origin in $(n + 1)$ space, and the input vector was chosen from a uniform distribution over the 2^n binary-valued input vectors. The weight vector was then perturbed in a random direction by the desired amount and the desired number of errors were introduced in the input vector. The output category to which the reference weight vector mapped the reference input vector was compared with the output category to which the perturbed weight vector mapped the perturbed input vector to determine whether or not the perturbations resulted in an Adaline decision error. This procedure was repeated 18 000 times for each data point.

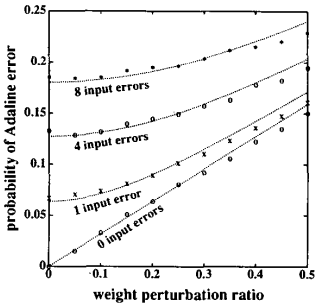


Fig. 6. Probability of Adaline decision error versus weight perturbation ratio for various input perturbation ratios. The results are shown for Adalines with 99 variable inputs. The continuous curves represent the probability of error as given by (36) and the data points depict the experimental frequency of error. For Adalines with 99 variable inputs, one input error corresponds to an input perturbation ratio of 0.2, four input errors to an input perturbation ratio of 0.4, and eight input errors to an input perturbation ratio of 0.57.

Data was generated for Adalines with various numbers of inputs. Fig. 6 shows a comparison of the computer generated data and the theoretical results for Adalines with 99 variable units. The continuous curves represent the theoretical results whereas the data points depict the computer-generated experimental results. Each curve shows the probability of an Adaline decision error as a function of weight perturbation ratio for a specific input perturbation ratio. The input perturbation ratio associated with each error curve is given in terms of the number of input errors. The input perturbation ratio for an Adaline with n variable inputs and h input errors is given by (27). For an Adaline with 99 variable inputs, one input error corresponds to an input perturbation ratio of 0.2, four input errors correspond to an input perturbation ratio of 0.4, and eight input errors correspond to an input perturbation ratio of 0.57.

The approximation for the probability of an Adaline error (36) assumes a large number of inputs as well as small weight and input perturbation ratios. From Fig. 6, we see that the agreement between theoretical and experimental results is good for weight and input perturbation ratios less than 0.5. Perturbation ratios in this range are most important from a practical standpoint. Although the re-

sults shown in Fig. 6 apply to Adalines with 99 variable inputs, we found good agreement between theoretical and experimental results for Adalines with as few as 9 variable inputs.

IX. Probability of Madaline Output Error Due to Weight Errors

Our final goal is the determination of the sensitivity of a network of Adalines to changes in the weights. For this purpose, consider a fixed Madaline network with arbitrarily chosen weights as a reference network. A perturbed network is generated from the reference network by adding randomly generated perturbation vectors (of desired magnitude) to the original weight vectors associated with each of the Adalines in the network. The magnitudes of the random perturbation vectors are chosen so that all Adalines have the same weight perturbation ratio. In this section, we find the probability that a selected output of the perturbed network differs from the corresponding output of the reference network.

It is straightforward to compute the probability of error for a first-layer Adaline (i.e., the probability that the output of a first-layer Adaline in the perturbed network is different from the output of the corresponding first-layer Adaline in the original network). The same inputs are presented to both networks, so the only source of error in the first-layer Adalines is weight perturbation. According to (19), the probability of error for a first-layer Adaline PE_1 is approximately equal to the weight perturbation ratio divided by π:

$$PE_1 \approx \delta W/\pi. \tag{37}$$

The outputs of the first-layer Adalines serve as inputs to the second-layer Adalines. This means that the Adalines of the second layer and all subsequent layers are subject to input errors as well as weight perturbations. So for $l > 1$, the probability of error PE_l for an lth layer Adaline is a function of both the weight perturbation ratio for the network and of the input perturbation ratio for the lth layer. The number of input errors to the lth layer is the same as the number of output errors from layer $l-1$. Assuming that the weight vectors of the n_{l-1} Adalines on layer $l-1$ are independent, the probability that exactly k of these Adalines make output errors is computed using the binomial distribution with parameters n_{l-1} and PE_{l-1}:

$$\begin{aligned} &P(k \text{ output errors on layer } l-1) \\ &\quad = \left(n_{l-1}!/\left(k!\,(n_{l-1}-k)!\right)\right)(PE_{l-1})^k \\ &\qquad \cdot (1-PE_{l-1})^{n_{l-1}-k}. \end{aligned} \tag{38}$$

The probability of error for an Adaline on layer l of the network can be computed by conditioning the error event on the number of input errors to the Adaline and then appropriately summing the resulting conditional probabilities. For example, let $PE_{l|k}$ denote the conditional probability that an lth layer Adaline of the perturbed network makes an error given that k of its n_{l-1} inputs are in error. The input perturbation ratio resulting from the k input errors is $(4k/(n_{l-1}+1))^{1/2}$. Using (36), we find:

$$PE_{l|k} \approx (1/\pi)\left[4k/(n_{l-1}+1) + (\delta W)^2\right]^{1/2}. \tag{39}$$

The probability that a selected lth layer Adaline makes an error is found by weighting $PE_{l|k}$ by the probability of k input errors and summing over all possible values of k:

$$PE_l = \sum_{k=0}^{n_{l-1}} PE_{l|k} P(k \text{ output errors on layer } l-1) \tag{40}$$

where n_{l-1} is the number of Adalines on layer $(l-1)$.

For convenience, we will refer to (37)–(40) as the *binomial approximation* (this name is due to the use of the binomial distribution in (38)). Note that if we want to use the binomial approximation to estimate PE_L (the probability of error for a selected output of a perturbed network with L layers), we must first find PE_1, PE_2, $\cdots$, PE_{L-1}. An analytic expression for PE_L cannot be obtained. Numerical values of PE_L can be determined by computer, but this is a tedious process. For this reason, it seems appropriate to derive an easily computable approximation to PE_L.

We now present a slightly less rigorous approach for evaluating network sensitivity to weight changes which results in a simpler approximation for PE_L. As before, the probability that a first-layer Adaline makes an error is approximated by: $PE_1 \approx \delta W/\pi$. Given that there are n_1 Adalines on the first layer and assuming the weight vectors of the first-layer Adalines to be independent, the expected number of input errors to the second layer is $n_1 PE_1$. If we substitute $n_1 PE_1$ for the number of input errors h in (27), we arrive at the following approximation of the input perturbation ratio for layer 2:

$$\delta X_{\text{layer2}} \approx \left[4PE_1 n_1/(n_1+1)\right]^{1/2} \approx \left[4PE_1\right]^{1/2}. \tag{41}$$

The second approximation follows from the assumption that n_1 is large. The net perturbation ratio for the Adalines of the second layer is approximated by the square root of the sum of the squares of the input perturbation ratio and the weight perturbation ratio:

$$\begin{aligned} &\text{the net perturbation ratio for layer 2} \\ &\quad \approx \left[(\delta W)^2 + 4PE_1\right]^{1/2}. \end{aligned} \tag{42}$$

Recalling that $PE_2 \approx (1/\pi)$ (net perturbation ratio for layer 2) and substituting $\delta W/\pi$ for PE_1, we find the following approximation for PE_2:

$$PE_2 \approx (\delta W/\pi)\left(1 + 4/(\pi \delta W)\right)^{1/2}. \tag{43}$$

This sequence of approximations can be repeated to find PE_3 from PE_2, then again to find PE_4 from PE_3, $\cdots$, and so on. Propagating the probability of error from one layer to the next in this manner, it is found that PE_l, the probability that the output of a lth layer Adaline is in error is approximately:

$$\begin{aligned} PE_l \approx \delta W/\pi(1 + \beta(1 + \beta(\cdots(1 \\ + \beta(1+\beta)^{1/2})^{1/2}\cdots)^{1/2})^{1/2})^{1/2} \end{aligned} \tag{44}$$

where

$$\beta \triangleq 4/(\pi\delta W) \tag{45}$$

and the number of square roots in the approximation for PE_l is $l - 1$. We will refer to (44) as the *square root approximation*.

In comparison to the binomial approximation, the square root approximation is easy to evaluate and is independent of the number of Adalines per layer. In order to illustrate the dependence of the binomial approximation on the number of Adalines per layer, we have used the binomial approximation to compute PE_l for networks which have equal numbers of Adalines on each layer. An n input-per-Adaline Madaline is one with n inputs, n first-layer Adalines, n second-layer Adalines, etc. Fig. 7 compares the binomial approximation for networks with 99, 199, and 499 Adalines per layer to the square root approximation which is independent of the number of layers. It is interesting to note that as the number of Adalines per layer increases, the binomial approximation becomes less and less dependent on the number of Adalines per layer. In fact, as the number of Adalines per layer increases, the probability of error as predicted by the binomial approximation approaches the probability of error predicted by the square root approximation. This comparison is shown for one-, two-, three-, and four-layer networks. The difference between the probabilities of error for networks with 499 Adalines per layer and 99 Adalines per layer is more pronounced for small weight perturbation ratios and for networks with many layers. Since the first-layer Adalines are not subject to input errors, the probability of error for the one-layer networks is independent of the number of inputs per Adaline.

The main difference in the derivations of the square root and binomial approximations is the manner in which the probability of error is propagated from one layer to the next. In the derivation of the binomial approximation, we conditioned the error event for an lth layer Adaline on all possible numbers of input errors to the Adaline. The conditional probabilities were weighted by the probabilities of the events on which they were conditioned and then summed. In the derivation of the square root approximation, we neglected the computation of the conditional probabilities and instead calculated the probability of error for the lth layer Adalines based on the expected number of input errors to the lth layer.

In summary, we have derived two approximations for the probability of a Madaline output error as a result of weight errors. Both approximations were derived by propagating the probability of error through the network from one layer to the next. The binomial approximation (37)–(40) is the most accurate but requires evaluation by computer and is dependent on the number of Adalines per layer. As the number of Adalines per layer increases, the probability of error approaches a limit and is essentially independent of the number of Adalines per layer. Fig. 7 suggests that this limit is given by the square root approximation (44). The square root approximation is not only much easier to use, but agrees very closely with the binomial approximation for large numbers of Adalines per layer.

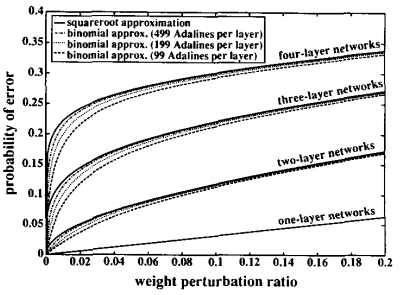


Fig. 7. Comparison of the square root and binomial approximations for the probability of error for a single Madaline output as a function of weight perturbation ratio. The binomial approximation is shown for networks for 99, 199, and 499 Adalines per layer. The square root approximation is independent of the number of Adalines per layer.

In the next section, we compare the probability of a Madaline output error as predicted by the square root approximation to the experimental frequency of Madaline error as found by computer simulation.

X. Simulation Results

A computer simulation was run to obtain experimental results for comparison with the theoretical results of the previous section. The purpose of the simulation was to find the experimental frequency of error for a Madaline output as a function of the weight perturbation ratio. To do so, a randomly generated weight vector was assigned to each Adaline of the network. A perturbed network was then generated from this (reference) network by perturbing each of the weight vectors in some random direction by the desired amount. A randomly selected input vector was then presented to both the reference network and the perturbed network and the outputs of the reference and perturbed networks were compared. Each data point is based on over 4000 comparisons.

The results of the simulation are contrasted against the probability of error as predicted by the square root approximation (44) in Fig. 8. Fig. 8(a) shows the results for networks with 49 Adalines per layer and Fig. 8(b) shows the results for networks with 299 Adalines per layer. In both cases, results are shown for one-, two-, three-, and four-layer networks. The same four theoretical curves are drawn in each of the figures since the square root approximation is independent of the number of inputs per Adaline. Comparing Fig. 8(a) and (b), we see that for the small weight perturbation ratios, the networks with 299 Adalines per layer have a slightly higher experimental probability of error than the networks with 49 Adalines per layer. However, for weight perturbation ratios greater than 5 percent, the difference is negligible. It is interesting to note that the agreement between theoretical and experimental results hold for weight perturbation ratios as high as 50 percent.

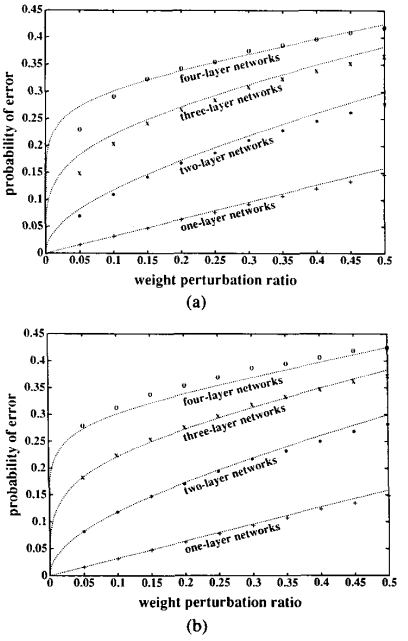


Fig. 8. Probability of error for a single Madaline output as a function of the weight perturbation ratio. Theoretical results (shown by the continuous curves) are based on the square root approximation. Data points depict the experimental frequency of error for a chosen output. (a) Results for Madalines with 49 Adalines per layer. (b) Results for Madalines with 299 Adalines per layer.

XI. Conclusion

In this paper, we have investigated the sensitivity of the input–output mapping of a feedforward layered network of Adaline elements (a Madaline) to errors in the weights. We began by analyzing the sensitivity of the Adaline to both input errors and weight errors. It was found that the probability of an Adaline decision error due to a combination of input errors and weight errors can be approximated as:

$$P(\text{Adaline Decision Error}) \approx (1/\pi)\left[(\delta X)^2 + (\delta W)^2\right]^{1/2} \tag{46}$$

where δX and δW are the input and weight perturbation ratios, respectively. In deriving this approximation, we assumed small weight and input perturbation ratios as well as a large number of Adaline inputs. It was found that agreement between theoretical and experimental results was good for weight and input perturbation ratios less than 0.5 and for Adalines with as few as 9 inputs.

The results on Adaline sensitivity were then applied to feedforward networks of Adaline elements. Two approximations were derived which predict the probability of error for a single output of the network as a function of the percentage error in the weights. The binomial approximation (37)–(40) is the most accurate but requires evaluation by computer and depends on the network size (the number of layers and the number of Adalines per layer). The square root approximation (44) is much easier to evaluate and is independent of the number of Adalines per layer. However, it is based on the assumption that this number is large. The square root approximation is repeated below for convenience:

$$\mathrm{PE}_L \approx \delta W/\pi(1 + \beta(1 + \beta(\cdots(1 + \beta(1 + \beta)^{1/2})^{1/2}\cdots)^{1/2})^{1/2})^{1/2} \tag{47}$$

where

$\mathrm{PE}_L \triangleq$ probability of error for a single output of an L-layer network with weight perturbation ratio δW,

$\delta W \triangleq$ weight perturbation ratio $\triangleq |\Delta W|/|W|$,

$\beta \triangleq 4/(\pi \delta W)$,

and the number of square roots in the approximation for PE_L is $L - 1$.

In Fig. 7, we compared the probability of error curves resulting from the binomial and square root approximations. Based on this comparison, we concluded that as the number of Adalines per layer increases (while the weight perturbation ratio and the number of layers remains constant), the probability of error as predicted by the binomial approximation approaches the probability of error predicted by the square root approximation. As long as there are enough Adalines per layer so that the difference between the probabilities of error predicted by the two approximations is "negligible," the probability of error is essentially independent of the number of Adalines per layer. The minimum number of Adalines per layer required for this independence assumption to hold depends on the weight perturbation ratio and on the number of layers in the network. The smaller the weight perturbation ratio and the greater the number of layers, the greater the number of Adalines per layer there must be before the probability of error becomes "independent" of the number of Adalines per layer.

A computer simulation was run to find the experimental frequency of error for an output of a Madaline with weight errors. A comparison of the experimental and theoretical data was illustrated in Fig. 8. Provided the number of Adalines per layer was sufficiently large (on the order of 100 or more), the agreement between the theoretical and experimental data was excellent. Good agreement was obtained even when the number of Adalines per layer was as small as 49. The agreement between theoretical and experimental results was found to hold for weight perturbation ratios as high as 50 percent.

Acknowledgment

The authors wish to thank Dr. H. Rauch and the reviewers of this paper for many valuable suggestions on the preparation of the manuscript.

References

[1] R. G. Winter and B. Widrow, "Madaline Rule II: A training algorithm for neural networks," in *Proc. IEEE Second Int. Conf. Neural Networks*, vol. I. (San Diego, CA), July 1988, pp. 401–408.

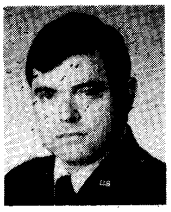
[2] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing*. Cambridge, MA: M.I.T. Press, 1986.
[3] B. Widrow, "Generalization and information storage in networks of Adaline 'neurons'," in *Self-Organizing Systems 1962*, M. C. Yovitz, G. T. Jacobi, and G. D. Goldstein, Eds. Washington, DC: Spartan Books, 1962, pp. 435-461.
[4] B. Widrow and R. G. Winter, "Neural nets for adaptive filtering and adaptive pattern recognition," *IEEE Computer Mag.*, pp. 25-39, Mar. 1988.
[5] B. Widrow and M. E. Hoff, Jr., "Adaptive switching circuits," in *IRE WESCON Convention Record*, pt. 4, Sept. 1960, pp. 96-104.
[6] D. M. Y. Sommerville, *An Introduction to the Geometry of N Dimensions*. London, England: Methuen and Co., 1929.
[7] M. E. Hoff, Jr., "Learning phenomena in networks of adaptive switching circuits," Ph.D. dissertation, Dept. Electrical Eng., Stanford Univ., Stanford, CA, June 1962.
[8] F. H. Glanz, "Statistical extrapolation in certain adaptive pattern-recognition systems," Ph.D. dissertation, Dept. Electrical Eng., Stanford Univ., Stanford, CA, May 1965.
[9] R. G. Winter, "Madaline rule II: A new method for training networks of Adalines," Ph.D. dissertation, Dept. Electrical Eng., Stanford Univ., Stanford, CA, Jan. 1989.

*

Maryhelen Stevenson (S'89) studied electrical engineering at the Institut National Polytechnique de Toulouse in France in 1981-1982, she received the B.E.E. degree from the Georgia Institute of Technology in 1983, and the M.S.E.E. degree from Stanford University in 1984. She is currently studying toward the Ph.D. degree in the Electrical Engineering Department at Stanford University.

She has been an employee of the Hughes Aircraft Company since 1978. Her research interests include adaptive signal processing and neural networks.

Ms. Stevenson is a student member of INNS as well as a member of Pi Kappa Phi, Eta Kappa Nu, and Tau Beta Pi.



Rodney Winter (S'87-M'88) received the B.S.E.E. and M.S.E.E. degrees from Purdue University in 1977. He received the Ph.D. degree from Stanford University in 1989 through the Air Force Institute of Technology.

He is a Major in the United States Air Force. He is currently assigned to the 20th Tactical Fighter Wing at RAF Upper Heyford, U.K., where he flies the F-111 aircraft. He also continues to research the application of neural networks to signal processing and pattern recognition problems.

Dr. Winter is a member of Eta Kappa Nu.

*

Bernard Widrow (M'58-SM'75-F'76) received the S.B., S.M., and Sc.D. degrees from M.I.T. in 1951, 1953, and 1956, respectively.

He is a Professor of Electrical Engineering at Stanford University. Before joining the Stanford faculty in 1959, he was with the Massachusetts Institute of Technology, Cambridge, MA. He is presently engaged in research and teaching in neural networks, pattern recognition, adaptive filtering, and adaptive control systems. He is Associate Editor of the journals *Adaptive Control and Signal Processing*, *Neural Networks*, *Information Sciences*, and *Pattern Recognition*, and is also coauthor with S. D. Stearns of *Adaptive Signal Processing* (Prentice-Hall).

Dr. Widrow is a member of the American Association of University Professors, the Pattern Recognition Society, Sigma Xi, and Tau Beta Pi. He is a fellow of the American Association for the Advancement of Science. He is President of the International Neural Network Society. He received the IEEE Alexander Graham Bell Medal in 1986 for exceptional contributions to the advancement of telecommunications.