901164

AD-A257 306

NWC TP 7191

2

Low Sensitivity Interpolation Using Feed-Forward Neural Networks With One Hidden Layer

by
Jorge M. Martin
Research Department

DECEMBER 1991

NAVAL WEAPONS CENTER
CHINA LAKE, CA 93555-6001

Approved for public release; distribution is unlimited.

Naval Weapons Center

FOREWORD

A technique and an algorithm for direct determination of weights for interpolation with low sensitivity using a neural net with one hidden layer are documented. These are results from a research project jointly sponsored by the Office of Naval Research and the Independent Research Program of the Naval Weapons Center, where the work was done. The work was done October 1990 to October 1991 under Program Element 0601153N, Research Project RR014-05-OK, RR052-02-02, R&T Project Code 411p 002---03, Type of Institution 12, and Program Element 61152N, Task Area RR00NW, Work Unit 13807004.

This report has been reviewed for technical accuracy by W. O. Alltop.

Approved by
R. L. DERR, *Head*
Research Department
27 December 1991

Under authority of
D. W. COOK
Capt., U.S. Navy
Commander

Released for publication by
W. B. Porter
Technical Director

NWC TECHNICAL PUBLICATION 7191

Published by . . . . Technical Information Department
Collation . . . . Cover, 20 leaves
First printing . . . . 86 copies

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (*Leave blank*)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	December 1991	Final, October 90 - October 91

4. TITLE AND SUBTITLE Low Sensitivity Interpolation Using Feed-Forward Neural Networks with one Hidden Layer	5. FUNDING NUMBERS PE 0601153N RP RR014-05-OK RR 052-02-02 TA RR 00NW WV 13807004
6. AUTHOR(S) Jorge M. Martin	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Weapons Center China Lake, CA 93555-6001	8. PERFORMING ORGANIZATION REPORT NUMBER NWC TP 7191
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Office of Naval Research/Code 1111 800 N. Quincy Street Arlington, VA 22217	10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT	12b. DISTRIBUTION CODE
Approved for public release; distribution unlimited	

13. ABSTRACT (*Maximum 200 words*)

(U) It is possible to assign directly the weights of a feed-forward neural net with one hidden layer so that the network interpolates through a given set of input-output points exactly and in such a way that the sensitivity to noise at the points of interpolation is as small as desired. This is demonstrated with a constructive proof. The weight assignment for exact interpolation requires the inversion of a nonsingular matrix. If the exact interpolation requirement is relaxed, then the inversion of that matrix can be avoided. It is possible to determine weights so that the network approximately interpolates through the set of points with any desired degree of accuracy and with a sensitivity as small as desired. Both the accuracy of interpolation and the sensitivity to noise are controlled by the size of the weights in the first layer of weights. Estimates on how large these weights have to be to achieve a desired interpolation accuracy and noise sensitivity are derived. An algorithm for approximate interpolation with low sensitivity is presented and illustrated with simple examples.

14. SUBJECT TERMS	15. NUMBER OF PAGES
weight assignment, exact/approximate interpolation, low sensitivity, total derivative	39
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

CONTENTS

ACKNOWLEDGMENT

The author is grateful to William Alltop for numerous helpful discussions.

1. INTRODUCTION

This paper considers the assignment of weights for neural networks with one hidden layer so that the network interpolates through a given finite set of input-output points with low sensitivity to noise in the input patterns. The sensitivity at the input patterns is minimized by minimizing the derivative of the input-output map of the interpolating network at the input patterns. The idea is that if the derivative at an input point is small, then a small variation around that point will produce a small variation in the output. This idea is made precise in the next section.

The approach presented here gives a direct method for determining the weights. The input-output map defined by these weights interpolates through the given set of points **exactly** and the derivative at the input patterns can be made arbitrarily small. The inversion of a nonsingular matrix is required for exact interpolation. If the exact interpolation requirement is relaxed, then the inversion of that matrix can be circumvented. It is possible to determine weights so that the network approximately interpolates through the given set of points with any desired degree of accuracy and with a sensitivity as small as desired. Both the accuracy of interpolation and the sensitivity to noise are controlled by the size of the weights in the first layer of weights. Estimates on how large these weights have to be to achieve a desired interpolation accuracy and noise sensitivity are also presented, as well as an algorithm for determining the weights.

Other authors have studied direct methods for weight assignment. By direct methods we mean nonrecursive methods; that is, methods that determine the weights as a well-defined, explicit function of the input-output pairs to be implemented. In Reference 1 it is shown how to approximately interpolate, with any desired degree of accuracy, through 2m-1 points with a network that has m neurons in the hidden layer and sigmoidal activation functions. It is also known that one can exactly interpolate through m+1 points with a network that has m neurons in the hidden layer and different types of activation functions (see for instance References 2 through 4). Here, the interpolation is through m+1 points using a network with m neurons in the hidden layer. The interpolation is done in such a way that the derivative at the points of interpolation can be controlled and can be made arbitrarily small. The input and output spaces can be multidimensional.

The weight assignment techniques for approximate interpolation can be applied to find a good set of initial weights for problems that involve learning more points than the degrees of freedom of the net. This can be an important application, since the speed of convergence of iterative learning algorithms is well known to depend severely on the choice of initial weights (References 5 and 6).

The notation required to present these results is introduced in Section 2, where we also state one of our main results and some preliminary results. In Section 3, we define weights and biases for a family of neural networks that solve the exact interpolation problem. The family is parametrized by a vector $w \in \mathbf{R}^m$ whose size controls the derivative of the input-output map at the interpolation points. We show that these derivatives can be made arbitrarily small by increasing the components of the vector w. Approximate interpolation with small sensitivity is addressed in Section 4. An algorithm for approximate interpolation with low sensitivity is presented and illustrated with simple examples. Some of the more tedious proofs are relegated to the Appendix.

2. NOTATION AND STATEMENT OF MAIN RESULTS

We shall consider feed-forward neural networks with one hidden layer consisting of m neurons, each of which has a nonlinear activation function that will be denoted by S. The activation function S is assumed to be a continuous function mapping the real line $\mathbf{R}$ into the open interval (-1, 1) with $\lim_{t \to \pm\infty} S(t) = \pm 1$.

For an m-vector y with components $y_1, y_2, \ldots, y_m$, we can define the m-dimensional sigmoid S_m by the formula

$$S_m(y) = \begin{bmatrix} S(y_1) \\ S(y_2) \\ \cdot \\ \cdot \\ \cdot \\ S(y_m) \end{bmatrix} \qquad (y \in \mathbf{R}^m) \ .$$

The collection of k by ℓ real matrices is denoted by $\mathbf{R}^{k \times \ell}$, and the space of k-dimensional real vectors is denoted by $\mathbf{R}^k$, where k and ℓ are any two positive integers. If the network has n inputs and ℓ outputs, then the transfer function (input-output map) of the network is a function $F : \mathbf{R}^n \to \mathbf{R}^\ell$ of the form

$$F(Z) = \alpha_o + \alpha \, S_m \, (WZ + \beta) \qquad (Z \in \mathbf{R}^n)$$

where $W \in \mathbf{R}^{m \times n}$ represents the first layer of weights, $\alpha \in \mathbf{R}^{\ell \times m}$ represents the second layer of weights, and the vectors $\beta \in \mathbf{R}^m$ and $\alpha_0 \in \mathbf{R}^\ell$ are bias vectors.

If S is differentiable, then S_m is differentiable as well as F. Let S'_m and F' denote the derivatives of S_m and F, respectively. Then $S'_m : \mathbf{R}^m \to \mathbf{R}^{m \times m}$ and $F' : \mathbf{R}^n \to \mathbf{R}^{\ell \times n}$ are given by

$$S'_m(y) = \mathrm{diag}\,(S'(y_1), S'(y_2), \ldots, S'(y_m)) \qquad (y \in \mathbf{R}^m)$$

$$F'(Z) = \alpha\, S'_m(WZ + \beta)W \qquad (Z \in \mathbf{R}^n) \quad , \tag{2.1}$$

where S' denotes the derivative of S, and $\mathrm{diag}(S'(y_1), \ldots, S'(y_m))$ is a diagonal matrix with $S'(y_1)$, $S'(y_2)$, ..., $S'(y_m)$ along the diagonal. Note that the ij^{th} component of the ℓ x n-matrix $F'(Z)$ is given by $[F'(Z)]_{ij} = \frac{\partial F_i}{\partial Z_j}(Z)$, where F_i is the i^{th} component of F and Z_j is the j^{th} component of Z $(1 \leq i \leq \ell,\ 1 \leq j \leq n)$.

Remark 2.1. For a vector valued function $F : \mathbf{R}^n \to \mathbf{R}^\ell$ such as the one above, the (total) derivative $F'(Z)$ of F at $Z \in \mathbf{R}^n$ is, by definition (see Reference 7 or 8), a linear transformation from $\mathbf{R}^n$ to $\mathbf{R}^\ell$ satisfying

$$\lim_{h \to 0} \frac{\| F(Z+h) - F(Z) - F'(Z)h \|}{\| h \|} = 0$$

where $\| \cdot \|$ denotes the underlying vector norm. This means that for every $\varepsilon > 0$ there exists a $\delta > 0$ such that if $\| h \| < \delta$, then

$$\| F(Z+h) - F(Z) - F'(Z)\, h \| \leq \varepsilon \|h\| \ .$$

For $\| h \| < \delta$, the above inequality implies

$$\| F(Z+h) - F(Z) \| \leq \| F'(Z)\, h \| + \varepsilon \| h \| < [\| F'(Z) \| + \varepsilon] \| h \| \ . \tag{2.2}$$

Inequality 2.2 has a significant interpretation. If Z represents a fixed input to the network with desired output F(Z) and h represents a small ($\| h \| < \delta$) perturbation to the exact input Z, then Inequality 2.2 asserts that the output $F(Z + h)$ to the perturbed input $(Z + h)$ will be within a distance $\delta[\| F'(Z) \| + \varepsilon]$ of the desired output F(Z). Therefore, by making $\| F'(Z) \|$ small, the output of the network to an input perturbed by noise will remain close to the desired output.

////*

Given a finite set of input-output pairs,

$$\Omega \equiv \{(x_i,y_i) \in \mathbf{R}^n \times \mathbf{R}^\ell : 0 \leq i \leq m \text{ and } x_i \neq x_j \text{ when } i \neq j\} \quad ,$$

we shall say that F *interpolates through* Ω if $F(x_i) = y_i$ for $i = 0, 1, 2, ..., m$.

Throughout this paper, the matrix W of first layer of weights will be given by an outer product $W = wv$, where $v^T \in \mathbf{R}^n$ will be fixed and chosen so that $vx_0 < vx_1 < ... < vx_m$. The m-vector w will belong to an unbounded open subset G of $\mathbf{R}^m$. The second layer of weights matrix α and the bias vectors α_0 and β will be defined as functions on G. Thus, we shall define (in the next section) functions $\alpha : G \rightarrow \mathbf{R}^{\ell \times m}$, $\beta : G \rightarrow \mathbf{R}^m$, and $\alpha_0 : G \rightarrow \mathbf{R}^\ell$, and a family of neural networks F_w ($w \in G$) of the form

$$F_w(z) = \alpha_0(w) + \alpha(w)\, S_m(wvz + \beta(w)) \qquad (z \in \mathbf{R}^n) \tag{2.3}$$

such that F_w interpolates through Ω for every $w \in G$. Moreover, under certain conditions on the sigmoid S, the function $\beta : G \rightarrow \mathbf{R}^m$ can be chosen so that

$$\lim_{w \to \infty} F_w'(x_i) = 0 \qquad \text{for } 0 \leq i \leq m \quad .$$

The notation $w \rightarrow \infty$ means that $w_i \rightarrow \infty$ for all $i = 1, 2, ..., m$, where w_i ($1 \leq i \leq m$) are the components of $w \in \mathbf{R}^m$.

These results, when combined with Remark 2.1, show that there exist neural networks that interpolate through the set Ω with an arbitrarily small sensitivity to noise at the inputs x_i ($0 \leq i \leq m$).

---

* The symbol //// indicates the end of a proof, an example, or a remark.

3. WEIGHT ASSIGNMENT FOR EXACT INTERPOLATION AND DERIVATIVE CONTROL

In this section we shall define a set G in $\mathbf{R}^m$ and a family of neural networks $F_w : \mathbf{R}^n \to \mathbf{R}^\ell$ ($w \in G$) of the form

$$F_w(x) = \alpha_0(w) + \alpha(w)\, S_m(wvx + \beta(w)) \qquad (x \in \mathbf{R}^n) \tag{3.1}$$

such that F_w interpolates through Ω for every $w \in G$. To solve the interpolation problem, it is required only that the activation function $S : \mathbf{R} \to (-1, 1)$ be continuous with $\lim_{t\to\pm\infty} S(t) = \pm 1$.

Next we will show that if the sigmoid S satisfies certain conditions for derivative control, then the bias vector function $\beta : G \to \mathbf{R}^m$ can be defined in such a way that the derivative of F_w can be made arbitrarily small at the points of interpolation.

3.1. EXACT INTERPOLATION

Let a set of interpolation points $\Omega = \{(x_i, y_i) \in \mathbf{R}^n \times \mathbf{R}^\ell : 0 \le i \le m \text{ and } x_i \ne x_j$ when $i \ne j\}$ be given.

The first step is to find a vector $v^T \in \mathbf{R}^n$ such that

$$vx_0 < vx_1 < vx_2 < \ldots < vx_m \quad . \tag{3.2}$$

Such a vector v^T always exists as asserted by the next lemma; however, we may have to relable the x_i ($0 \le i \le m$).

Lemma 3.1. Given distinct points $x_0, x_1, x_2, \ldots, x_m$ in $\mathbf{R}^n$, there exists a vector $v^T \in \mathbf{R}$ such that $\{vx_i : 0 \le i \le m\}$ is a set of distinct numbers.

This lemma is proved in the Appendix.

Note that v denotes a row vector, while its transpose v^T denotes a column vector in $\mathbf{R}^n$.

Given $v^T \in \mathbf{R}^n$ satisfying Inequalities 3.2, we define $W = wv$ with $w \in \mathbf{R}^m$. Next, one selects any m continuous functions $A_k : (t_k, \infty) \to (0, \infty)$ $(1 \le k \le m)$ that **grow slower than linear**; that is, they satisfy the **Growth Condition**

$$\lim_{t \to \infty} [ta + A_k(t)] = \begin{cases} +\infty & \text{if } a \ge 0 \\ -\infty & \text{if } a < 0 \end{cases} \qquad (1 \le k \le m) \ . \tag{3.3}$$

For example, $A_k(t) = (t - t_k)^\varepsilon$ for $t > t_k$ and $0 < \varepsilon < 1$, $k = 1, 2, ..., m$.

The bias vector function β is defined on the open set $X \equiv \prod_{k=1}^{m} (t_k, \infty) \subset \mathbf{R}^m$. If $w \equiv [w_1, w_2, ..., w_m]^T \in X$, then the i^{th} component of $\beta(w)$ is given by

$$\beta_i(w) \equiv A_i(w_i) - w_i v x_i \qquad (w \in X, 1 \le i \le m) \ . \tag{3.4}$$

To simplify the notation, let $L_w : \mathbf{R}^n \to \mathbf{R}^m$ denote the affine transformation defined for each $w \in X$ by

$$L_w(x) \equiv wvx + \beta(w) \qquad (x \in \mathbf{R}^n) \ .$$

Note that for each $w \in X$, L_w is the transformation between the input layer and the hidden layer. By Equation 3.4, the i^{th} component of $L_w(x)$ can be written as

$$[L_w(x)]_i = w_i v (x - x_i) + A_i(w_i) \qquad (x \in \mathbf{R}^n, 1 \le i \le m) \ . \tag{3.5}$$

Let $\Delta(w)$ denote the m x m matrix whose j^{th} column equals

$$S_m(L_w(x_j)) - S_m(L_w(x_{j-1})) \qquad (1 \le j \le m, w \in X) \ . \tag{3.6}$$

Note that since the sigmoid S and the functions A_i $(1 \le i \le m)$ are continuous, the matrix valued mapping $w \to \Delta(w)$ defines a continuous map $\Delta : X \to \mathbf{R}^{m \times m}$.

The set G is defined to be the collection of all vectors $w \in X$ for which $\Delta(w)$ is an invertible matrix. We shall see shortly that G is an unbounded open set in $\mathbf{R}^m$.

The matrix valued function $\alpha : G \to \mathbf{R}^{\ell \times m}$ is defined by

$$\alpha(w) \equiv Y\, \Delta^{-1}(w) \quad (w \in G) \ , \tag{3.7}$$

where $Y \equiv [y_1 - y_0 \vdots y_2 - y_1 \vdots \ldots \vdots y_m - y_{m-1}] \in \mathbf{R}^{\ell \times m}$.

Finally, $\alpha_0 : G \to \mathbf{R}^{\ell}$ is defined as

$$\alpha_o(w) \equiv y_o - \alpha(w)\, S_m\,(L_w(x_o)) \qquad (w \in G) \ . \tag{3.8}$$

Our first theorem shows how this construction solves the interpolation problem. It also shows why it suffices to have m neurons in the hidden layer to interpolate through (m+1) points.

Theorem 3.1. For each $w \in G$, the layered neural network

$$F_w(x) \equiv \alpha_o(w) + \alpha(w)\, S_m(L_w(x)) \qquad (x \in \mathbf{R}^n)$$

interpolates through Ω.

Proof. The proof is by induction. Fix $w \in G$. The definition of $\alpha_0(w)$ (Equation 3.8) clearly implies $F_w(x_0) = y_0$. Assume that $F_w(x_k) = y_k$ for $0 \leq k < m$. If e_{k+1} denotes the $(k+1)^{st}$ column of the $m \times m$ identity matrix, then

$$\begin{aligned} F_w(x_{k+1}) - F_w(x_k) &= Y\, \Delta^{-1}(w)\, [S_m(L_w(x_{k+1})) - S_m(L_w(x_k))] \\ &= Y\, e_{k+1} = y_{k+1} - y_k \ . \end{aligned}$$

Here we used the definition of $\alpha(w)$ (Equation 3.7) and the definition of the $(k+1)^{st}$ column of $\Delta(w)$ (see Expression 3.6). It follows that $F_w(x_{k+1}) = y_{k+1}$. This completes the proof. ////

For the construction above to work, it is essential that $\Delta(w)$ be invertible for some values of w. This is guaranteed by the next proposition, which is a consequence of the Growth Condition 3.3 and the asymptotic properties of S.

Proposition 3.1. $\lim_{w \to \infty} \Delta(w) = 2I_m$, where I_m denotes the m x m identity matrix. Consequently, G is an unbounded open subset of $\mathbf{R}^m$. More precisely, there exist $T_k \geq t_k$ large enough ($1 \leq k \leq m$) such that the product $\prod_{k=1}^{m} (T_k, \infty)$ is contained in G.

Recall that $\lim_{w \to \infty}$ means that $w_i \to \infty$ for all i = 1, 2, ..., m.

Proof. If U denotes the collection of all invertible matrices in $\mathbf{R}^{m \times m}$, then U is an open set containing $2I_m$. Therefore, if the above limit holds, then U contains $\Delta(w)$ for all w large enough. This implies that G contains the product $\prod_{k=1}^{m} (T_k, \infty)$ for T_k large enough ($1 \leq k \leq m$). Moreover, since $\Delta : X \to \mathbf{R}^{m \times m}$ is continuous, $G = \Delta^{-1}(U)$ is open in X, hence open in $\mathbf{R}^{m \times m}$.

To prove that $\Delta(w)$ converges to $2I_m$ for large w, let $\Delta_{ij}(w)$ denote the ij^{th} entry of $\Delta(w)$, $1 \leq i \leq m$, $1 \leq j \leq m$. Equation 3.5 and Expression 3.6 give

$$\Delta_{ij}(w) = S(w_i v(x_j - x_i) + A_i(w_i)) - S(w_i v(x_{j-1} - x_i) + A_i(w_i))$$
$$= \begin{cases} S(A_i(w_i)) - S(w_i v(x_{i-1} - x_i) + A_i(w_i)) & \text{if } j = i \\ S(w_i v(x_{i+1} - x_i) + A_i(w_i)) - S(A_i(w_i)) & \text{if } j = i+1 \\ S(w_i v(x_j - x_i) + A_i(w_i)) - S(w_i v(x_{j-1} - x_i) + A_i(w_i)) & \text{if } j < i \text{ or } j > i+1 \end{cases} .$$

Hence it follows from the choice of v (Inequalities 3.2), the asymptotic properties of S, and the Growth Condition 3.3 that

$$\lim_{w \to \infty} \Delta_{ij}(w) = \begin{cases} 2 \text{ if } j = i \\ 0 \text{ if } j \neq i \end{cases} .$$

This completes the proof of the proposition. ////

Remark 3.1. The Growth Condition 3.3 on the functions A_k $(1 \leq k \leq m)$ was instrumental in the proof of Proposition 3.1, which hinges on the fact that $\Delta(w)$ converges to an invertible matrix as $w \to \infty$, and this guarantees the invertibility of $\Delta(w)$ for w in the unbounded set G. It should be pointed out that if one is only interested in solving the interpolation problem, then one may do without the Growth Condition 3.3 and replace the functions A_k by arbitrary constants. If A_k are constants $(1 \leq k \leq m)$, then $\Delta(w)$ still converges to an invertible matrix M as $w \to \infty$. Thus, Proposition 3.1 will hold for arbitrary constants A_k if $2I_m$ is replaced by the matrix M, which has the form

$$M = \begin{bmatrix} a_1 & b_1 & 0 & \cdots & 0 \\ 0 & a_2 & b_2 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & & \\ 0 & 0 & \cdots & a_{m-1} & b_{m-1} \\ 0 & 0 & \cdots & 0 & a_m \end{bmatrix}$$

with $a_k \equiv 1 + S(A_k)$ and $b_k \equiv 1 - S(A_k)$ $(1 \leq k \leq m)$. Note that $a_k \to 2$ and $b_k \to 0$ if $A_k \to \infty$ $(1 \leq k \leq m)$. Of course, Theorem 3.1 holds whenever $\Delta(w)$ is invertible, independent of what the limit of $\Delta(w)$ might be as $w \to \infty$. It is because the Growth Condition 3.3 will be required to control the derivative of F_w at x_i $(0 \leq i \leq m)$ that we chose to present this approach for solving the interpolation problem. Moreover, the fact that $\Delta^{-1}(w) \to \frac{1}{2} I_m$ as $w \to \infty$ will lead to a simple formula for the weight matrix α; namely, $\frac{1}{2} Y$, which will solve the interpolation problem approximately without matrix inversions. ////

3.2. DERIVATIVE CONTROL

To control the limiting behavior of F'_w at the points of interpolation as $w \to \infty$, we require that the functions A_k approach infinity as $w \to \infty$ in a particular way. Given $r_k \geq 0$, we assume that the functions $A_k : (t_k, \infty) \to (0, \infty)$ $(1 \leq k \leq m)$ in Equation 3.4 satisfy the Growth Condition 3.3 and the two conditions below:

$$\left.\begin{aligned} &\lim_{t \to \infty} t\, S'(A_k(t)) = 0 && \text{if } r_k = 0 \\ &t\, S'(A_k(t)) = r_k && \text{for } t > t_k \ \text{ if } r_k > 0 \end{aligned}\right\} \qquad (3.9)$$

$$\lim_{t \to \infty} t\, S'(at + A_k(t)) = 0 \quad \text{for } a \neq 0 \text{ and } 1 \leq k \leq m \quad . \tag{3.10}$$

These conditions for derivative control are satisfied by a vast class of differentiable sigmoids. This class includes commonly used sigmoids such as the hyperbolic tangent, for which

$$A_k(t) = \cosh^{-1}[\sqrt{t/r_k}\,] \quad , \quad t > r_k \quad , \quad \text{when } r_k > 0 \quad ,$$

and

$$A_k(t) = \cosh^{-1}[\sqrt{t^{1+\varepsilon}}] \quad , t > 1 \quad , \quad \text{for any } \varepsilon > 0 \text{ when } r_k = 0 \quad .$$

Another example is the inverse tangent $S(t) = \frac{2}{\pi} \tan^{-1}(t)$, $(t \in \mathbf{R})$, for which

$$A_k(t) = \sqrt{(2t/\pi r_k)-1} \quad , \quad t > \pi r_k/2 \quad , \quad \text{when } r_k > 0 \quad ,$$

and

$$A_k(t) = \sqrt{(2t^{3/2}/\pi)-1} \quad , \quad t > (\pi/2)^{2/3} \quad , \quad \text{when } r_k = 0 \quad .$$

For the logistic sigmoid,

$$S(t) = \frac{2}{\sqrt{\pi}} \int_{-\infty}^{t} e^{-x^2}\, dx - 1 \quad , \quad (t \in \mathbf{R})$$

$$A_k(t) = [\ln 2t - \ln \sqrt{\pi}\, r_k]^{1/2} \quad , \quad t > \sqrt{\pi}\, r_k/2 \quad , \quad \text{when } r_k > 0 \quad ,$$

and

$$A_k(t) = [\ln \frac{2}{\sqrt{\pi}}\, t^{1+\varepsilon}]^{1/2} \quad , \quad t > \sqrt[1+\varepsilon]{\sqrt{\pi}/2} \quad , \text{for any } \varepsilon > 0 \quad , \quad \text{when } r_k = 0 \quad .$$

It is not hard to show that the examples above satisfy Conditions 3.3, 3.9, and 3.10 (see Remark 3.2).

When the derivative of the sigmoid is strictly decreasing on some infinite interval of the positive real axis as in the examples above, then the derivative S' of the sigmoid is invertible on that interval. Thus, one may solve the equation $t\, S'(A_k(t)) = r_k$ to obtain a unique function A_k for each $r_k > 0$. If $(S')^{-1}$ denotes the inverse of S' on an appropriate domain, then $A_k(t) = (S')^{-1}(r_k/t)$ for appropriate values of t. When $r_k = 0$, one simply chooses a decaying function f such that $f(t) \to 0$ as $t \to \infty$ and solves $t\,S'(A_k(t)) = f(t)$ to get $A_k(t) = (S')^{-1}(\frac{1}{t}f(t))$ for t in an appropriate domain. For example, $f(t) = t^{-\varepsilon}$ $(t > 0)$, with $\varepsilon > 0$ judiciously chosen so that A_k satisfies the Growth Condition 3.3.

The following lemma sheds light on some relationships that exist among Conditions 3.3, 3.9, and 3.10 under certain assumptions on the sigmoid S.

Lemma 3.2. Suppose that $S : \mathbf{R} \to (-1, 1)$ is a differentiable odd function with S' nonincreasing on $(0, \infty)$. Suppose that for every $r_k \geq 0$ there exists $t_k > 0$ and $A_k : (t_k, \infty) \to (0, \infty)$ such that Condition 3.9 and Condition 3.3 with $a < 0$ hold. Then Condition 3.3 holds for $a = 0$ (and all $a > 0$) and Condition 3.10 holds for all $a \neq 0$.

The proof may be found in the Appendix.

Remark 3.2. It should be clear from Lemma 3.2 and the observations preceding it that for odd sigmoids with **strictly decreasing derivative** on some infinite interval of the positive real axis, the functions A_k satisfying Condition 3.9 always exist and are in fact unique when $r_k > 0$. Consequently, it is only the Growth Condition 3.3 with $a < 0$ that must be verified when dealing with such sigmoids. ////

Theorem 3.2. If the sigmoid S and the functions A_k $(1 \leq k \leq m)$ in the definition of β (Equation 3.4) satisfy Conditions 3.3, 3.9, and 3.10, then the family F_w $(w \in G)$ of Theorem 3.1 interpolates through Ω and

$$\lim_{w \to \infty} F'_w(x_k) = \begin{cases} 0 & \text{for } k = 0 \\ \\ \frac{1}{2} r_k (y_k - y_{k-1})\, v & \text{for } 1 \leq k \leq m \end{cases} .$$

Proof. Since A_k $(1 \leq k \leq m)$ satisfy the Growth Condition 3.3, it follows from Theorem 3.1 that F_w interpolates through Ω for all $w \in G$.

Now, by Equation 2.1 and the definitions of W, $\beta(w)$, L_w, and $\alpha(w)$, for each $w \in G$, $F'_w(x_k)$ is given by

$$F'_w(x_k) = Y\,\Delta^{-1}(w)\,S'_m(L_w(x_k))wv \qquad (0 \leq k \leq m) . \tag{3.11}$$

Since $\lim_{w\to\infty} \Delta^{-1}(w) = \frac{1}{2} I_m$ (Proposition 3.1), it suffices to show that

$$\lim_{w\to\infty} S'_m(L_w(x_k))\, w = \begin{cases} 0 & \text{for } k=0 \\ \\ r_k e_k & \text{for } 1 \le k \le m \end{cases} \quad , \tag{3.12}$$

where e_k denotes the k^{th} column of the mxm identity matrix I_m. Indeed, if the Limit 3.12 holds, then by Equation 3.11,

$$\lim_{w\to\infty} F'_w(x_k) = \begin{cases} \frac{1}{2} r_k\, Y\, e_k\, v = \frac{1}{2} r_k(y_k - y_{k-1})\, v & \text{for } 1 \le k \le m \\ \\ 0 & \text{for } k = 0 \quad . \end{cases}$$

To establish the Limit 3.12, consider the i^{th} component of the vector $S'_m(L_w(x_k))w$ $(1 \le i \le m,\ 0 \le k \le m)$

$$[S'_m(L_w(x_k))w]_i = w_i\, S'(w_i v(x_k - x_i) + A_i(w_i)) \quad . \tag{3.13}$$

When $i \ne k$, $v(x_k - x_i) \ne 0$. Thus, Equations 3.10 and 3.13 imply

$$\lim_{w_i\to\infty} [S'_m(L_w(x_k))w]_i = 0 \qquad (0 \le k \le m\ ,\ 1 \le i \le m\ ,\ i \ne k) \quad .$$

When $i = k$, the choice of A_k and Equations 3.9 and 3.13 imply

$$\lim_{w_k\to\infty} [S'_m(L_w(x_k))w]_k = r_k \qquad (1 \le k \le m) \quad .$$

The last two limits show that the Limit 3.12 holds. This completes the proof of the theorem. ////

By setting $r_k = 0$ for $k = 1, 2, \ldots, m$ in Theorem 3.2, we arrive at one of the main results of this paper.

Corollary 3.1. (Exact Interpolation With Low Sensitivity). There exists a family of neural networks $F_w : \mathbf{R}^n \rightarrow \mathbf{R}^\ell$ $(w \in G)$ such that each F_w interpolates through Ω and

$$\lim_{w \to \infty} F'_w(x_k) = 0 \quad \text{for} \quad 0 \le k \le m \ .$$

Proof. The examples appearing after Condition 3.10 give sigmoids S and functions A_k for every $r_k \geq 0$ that satisfy the hypothesis of Theorem 3.2. Thus, the corollary follows from the theorem with $r_k = 0$ $(1 \leq k \leq m)$. ////

Another result that follows as a special case of Theorem 3.2 when $n = \ell = 1$ states that the values of a one-input/one-output net with m hidden neurons can be exactly specified at m + 1 points and the derivatives at m of those points can be approximately assigned with any degree of accuracy except for a sign restriction.

Theorem 3.3. Let $(x_k, y_k) \in \mathbf{R}^2$ $(0 \leq k \leq m)$ be m + 1 points such that $x_0 < x_1 < \ldots < x_m$ and let d_k $(0 \leq k \leq m)$ be m + 1 real numbers satisfying one of the two conditions below:

(a) $d_0 = 0$ and for $k > 0$ $d_k = 0$ if $y_k - y_{k-1} = 0$; otherwise, $d_k(y_k - y_{k-1}) \geq 0$.

(b) $d_m = 0$ and for $k < m$ $d_k = 0$ if $y_{k+1} - y_k = 0$; otherwise, $d_k(y_{k+1} - y_k) \geq 0$.

Then, there exists a family of neural nets $F_w : \mathbf{R} \rightarrow \mathbf{R}$ $(w \in G)$ with m hidden neurons such that

$$F_w(x_k) = y_k \quad \text{and} \quad \lim_{w \to \infty} F'_w(x_k) = d_k \quad (0 \le k \le m) \ .$$

Proof. Assume that Condition (a) holds. Set $r_k = 0$ if $y_k - y_{k-1} = 0$; otherwise, $r_k = \dfrac{2d_k}{y_k - y_{k-1}}$ $(1 \leq k \leq m)$. Let S denote any of the sigmoids in the examples appearing after Condition 3.10 and let A_k correspond to r_k $(1 \leq k \leq m)$ as in Theorem 3.2. Since $r_k \geq 0$ $(1 \leq k \leq m)$, Theorem 3.2 applies with $v = 1$ to yield the result.

If Condition (b) holds, set $x'_k = x_{m-k}$ and $y'_k = y_{m-k}$ $(0 \leq k \leq m)$ and let $r'_k = r_{m-k}$, where $r_k = 0$ if $y_{k+1} - y_k = 0$; otherwise , $r_k = \dfrac{2d_k}{y_{k+1}-y_k}$ $(0 \leq k \leq m-1)$. Since $r'_k \geq 0$ $(1 \leq k \leq m)$, Theorem 3.2 applies to the set $\Omega = \{(x'_k, y'_k) \in \mathbf{R}^2 : 0 \leq k \leq m\}$ with $v = -1$ and we obtain $F_w : \mathbf{R} \rightarrow \mathbf{R}$ $(w \in G)$ such that F_w interpolates through Ω and

$$\lim_{w \to \infty} F'_w(x'_k) = \begin{cases} 0 & \text{for } k = 0 \\ \frac{1}{2} r'_k (y'_k - y'_{k-1}) v & \text{for } 1 \leq k \leq m \end{cases} \quad . \tag{3.14}$$

Now, by Equation 3.14 and the definitions of x'_k, y'_k, and r'_k, we have

$$\begin{aligned} \lim_{w \to \infty} F'_w(x_k) &= \lim_{w \to \infty} F'_w(x'_{m-k}) = -\frac{1}{2} r'_{m-k}(y'_{m-k} - y'_{m-k-1}) \\ &= \frac{1}{2} r_k(y_{k+1} - y_k) = d_k \qquad (0 \leq k \leq m-1) \end{aligned}$$

and

$$\lim_{w \to \infty} F'_w(x_m) = \lim_{w \to \infty} F'_w(x'_0) = 0 = d_m \quad .$$

This completes the proof. ////

We close this section with some comments about the last result. A network with m hidden neurons, one input, and one output has 3m + 1 degrees of freedom; namely, the components of the m-vectors α^T, w, and β and the constant α_0. Theorem 3.3 exhibits a family F_w parametrized by vectors w belonging to the unbounded open set G in $\mathbf{R}^m$. Each F_w interpolates through m + 1 points. This accounts for m + 1 degrees of freedom. The parameter w accounts for m degrees of freedom. The remaining m degrees of freedom were utilized to approximately assign the derivatives at m of the interpolation points within the restrictions of Conditions (a) and (b) of Theorem 3.3.

4. APPROXIMATE INTERPOLATION WITH LOW SENSITIVITY, ALGORITHM, AND EXAMPLES

An interesting consequence of Proposition 3.1 is that $\lim_{w\to\infty} \Delta^{-1}(w) = \frac{1}{2} I_m$. Consequently, $\lim_{w\to\infty} \alpha(w) = \frac{1}{2} Y$. Thus, one may be tempted to replace the second matrix of weights $\alpha(w) = Y\,\Delta^{-1}(w)$ by $\frac{1}{2}$ Y, since this choice of weights avoids having to compute the inverse of $\Delta(w)$. With this choice of α the interpolation through Ω will not be exact. It will improve, however, as w increases. In this section, this idea will be explored. We shall derive conditions that determine how large w must be in order to approximately interpolate through Ω within a given error tolerance and with $[F'_w(x_k)]_{ij}$ within a prescribed distance from zero $(0 \le k \le m,\ 1 \le i \le \ell,\ 1 \le j \le n)$ using $\alpha = \frac{1}{2} Y$.

The neural network map with $\alpha = \frac{1}{2} Y$ will be denoted by T_w. It can be written as

$$T_w(x) = y_o + \frac{1}{2} Y\,[S_m(L_w(x)) - S_m(L_w(x_o))] \qquad (x \in \mathbf{R}^n) \ , \tag{4.1}$$

where L_w is defined in terms of w, v, and $\beta(w)$ as in Section 3 and w may be any vector in X.

Our first lemma gives a bound on the size of the error $T_w(x_j) - y_j$ in terms of the size of the vectors y_j $(0 \le j \le m)$. The absolute value of a real number z will be denoted by $|z|$ If z is a vector with components $z_1, z_2, \ldots, z_k$, then $|z| = [|z_1|, |z_2|, \ldots, |z_k|]^T$. If z' is another k-vector, then $|z| \le |z'|$ means $|z_i| \le |z'_i|$ for i = 1, 2, ..., k.

Remark 4.1. Since the sigmoid $S : \mathbf{R} \to (-1, 1)$ is continuous with $\lim_{t\to\pm\infty} S(t) = \pm 1$, given any number $\delta \in (0, 1)$ one can find $\alpha > 0$ large enough that $1 - \delta < S(t) < 1$ for all $t > \alpha$. ////

Lemma 4.1. Choose $\delta_1 \in (0, 1)$ and $\alpha > 0$ such that $1 - \delta_1 < S(t) < 1$ for all $t > \alpha$. If $w \in X$ has positive components and satisfies the following two conditions

$$A_\mu(w_\mu) > \alpha \qquad (1 \le \mu \le m) \tag{4.2}$$

$$w_\mu v\,(x_{\mu-1} - x_\mu) + A_\mu(w_\mu) < -\,\alpha \qquad (1 \le \mu \le m) \tag{4.3}$$

then

$$|T_w(x_j) - y_j| \le \frac{3}{2}\,\delta_1 \sum_{i=o}^{m} |y_i| \qquad (1 \le j \le m)\ . \tag{4.4}$$

This lemma is proved in the Appendix.

The next lemma gives a bound on the size of $[T_w'(x_k)]_{ij}$ in terms of the j^{th} component of v and the size of the i^{th} component of the differences $|y_\mu - y_{\mu-1}|$ $(1 \le \mu \le m,\ 1 \le i \le \ell,\ 1 \le j \le n)$.

Lemma 4.2. Choose $\delta_2 > 0$ and assume that S is an odd differentiable function with S' nonincreasing on $(0, \infty)$. If the neural network map is given by Equation 4.1 and if $w = [w_1, w_2, ..., w_m]^T \in X$ has positive components and satisfies the following two conditions,

$$0 < w_\mu S'(A_\mu(w_\mu)) < \delta_2 \qquad (1 \le \mu \le m) \tag{4.5}$$

$$\left.\begin{aligned} &0 < w_\mu S'(w_\mu v(x_{\mu-1} - x_\mu) + A_\mu(w_\mu)) < \delta_2 \quad \text{with} \\ &w_\mu v\,(x_{\mu-1} - x_\mu) + A_\mu(w_\mu) < 0 \qquad (1 \le \mu \le m) \end{aligned}\right\}\ , \tag{4.6}$$

then

$$|[T_w'(x_k)]_{ij}| \le \frac{\delta_2}{2}\,[\sum_{\mu=1}^{m} |(y_\mu - y_{\mu-1})_i\,|]\ |v_j|\,(0 \le k \le m,\ 1 \le i \le \ell,\ 1 \le j \le n), \tag{4.7}$$

where $(y_\mu - y_{\mu-1})_i$ denotes the i^{th} component of the vector $(y_\mu - y_{\mu-1})$, $\mu = 1, 2, ..., m$ and $[T_w'(x_k)]_{ij}$ is the ij^{th} entry of the matrix $T_w'(x_k)$.

The proof of this lemma may be found in the Appendix.

Lemma 4.1 establishes the connection between the asymptotic values of the sigmoid and the size of the errors in the approximate interpolation. It shows that the errors can be made arbitrarily small by choosing w large enough, as specified by Inequalities 4.2 and 4.3. Its counterpart, Lemma 4.2, shows the connection between the asymptotic values of the functions $f_{a,\mu}(t) \equiv t\, S'(at + A_\mu(t))$ $(t > t_\mu,\ 1 \leq \mu \leq m,\ a \in \mathbf{R})$ and the size of the derivatives at the points of interpolation. It shows that the derivatives can be made arbitrarily small by choosing w large enough as specified by Inequalities 4.5 and 4.6.

The next theorem, which is the main result of this section, puts together these results in a proof showing that when the functions $A_k : (t_k, \infty) \to (0, \infty)$ approach infinity as $w \to \infty$ in the particular way described in Section 3, then one can in fact find a vector $\boldsymbol{w} \in X$ such that the errors and the derivatives at the interpolation points are arbitrarily small for **all** $w \geq \boldsymbol{w}$. Before stating the theorem, it should be emphasized that the two lemmas above hold true if the functions $A_k : (t_k, \infty) \to (0, \infty)$ $(1 \leq k \leq m)$ are constant functions. That is, the values $A_\mu(w_\mu)$ appearing in the Inequalities 4.2, 4.3, 4.5, and 4.6 can be fixed constants independent of w_μ $(1 \leq \mu \leq m)$ without invalidating the proofs of the two lemmas. Notice, however, that Inequality 4.5 cannot hold for all $w \geq \boldsymbol{w}$ unless $S'(A_\mu(w_\mu))$ decreases as w_μ increases without a bound, forcing $A_\mu(w_\mu)$ to vary with w_μ. Since the main purpose of these two lemmas is to facilitate the proof of Theorem 4.1, which requires Inequality 4.5 to hold for all $w \geq \boldsymbol{w}$, we chose to state the lemmas in a manner that indicates the possibility that $A_\mu(w_\mu)$ may vary with w_μ $(1 \leq \mu \leq m)$.

Theorem 4.1. (Approximate Interpolation With Low Sensitivity). Assume that S is an odd differentiable function with S' nonincreasing on $(0, \infty)$. Let $A_k : (t_k, \infty) \to (0, \infty)$ satisfy the Growth Condition 3.3, Condition 3.10, and Condition 3.9 with $r_k = 0$ $(1 \leq k \leq m)$. Let $T_w : \mathbf{R}^n \to \mathbf{R}^\ell$ be given by Equation 4.1. Then, for any $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$, there exists $\boldsymbol{w} \in X$ such that for all $w \geq \boldsymbol{w}$,

$$|(T_w(x_j) - y_j)_i| < \varepsilon_1 \qquad (1 \leq i \leq \ell, 0 \leq j \leq m)$$

and

$$|[T'_w(x_k)]_{ij}| < \varepsilon_2 \qquad (1 \leq i \leq \ell, 1 \leq j \leq n, 0 \leq k \leq m) \ ,$$

where $(T_w(x_j) - y_j)_i$ denotes the i^{th} component of $T_w(x_j) - y_j$.

Proof. Let y_{ji} denote the i^{th} component of y_j $(1 \leq i \leq \ell, 0 \leq j \leq m)$. Let $\delta_1 \in (0, 1)$ and $\delta_2 > 0$ satisfy

$$\frac{3}{2}\,\delta_1 \sum_{j=0}^{m} |y_{ji}| \;<\; \varepsilon_1 \qquad (1 \le i \le \ell) \tag{4.8}$$

$$\frac{1}{2}\,\delta_2 \sum_{\mu=1}^{m} |(y_\mu - y_{\mu-1})_i|\; |v_j| \;<\; \varepsilon_2 \qquad (1 \le i \le \ell,\ 1 \le j \le n) \tag{4.9}$$

and choose $\alpha > 0$ as in Lemma 4.1. Then, clearly, by Lemmas 4.1 and 4.2, it suffices to show that there exists $\boldsymbol{w} \in X$ such that Inequalities 4.2, 4.3, 4.5, and 4.6 hold for all $w \ge \boldsymbol{w}$.

Fix $\mu \in \{1, 2, ..., m\}$. The Growth Condition 3.3 clearly implies that there exists $w'_\mu > t_\mu$ such that Inequalities 4.2 and 4.3 hold for all $w_\mu \ge w'_\mu$. Similarly, Condition 3.9 with $r_\mu = 0$ implies that there exists $w''_\mu > t_\mu$ such that Inequality 4.5 holds for all $w_\mu \ge w''_\mu$. Finally, the Growth Condition 3.3 and Condition 3.10 imply that there exists $w'''_\mu > t_\mu$ such that Inequality 4.6 holds for all $w_\mu \ge w'''_\mu$. By letting $\boldsymbol{w}_\mu = \max\{w'_\mu, w''_\mu, w'''_\mu\}$ for each $\mu \in \{1, 2, ..., m\}$, we obtain $\boldsymbol{w} = [\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_m]^T \in X$ with the required properties. ////

The functions A_μ $(1 \le \mu \le m)$ in Theorem 4.1 have in common that they all satisfy Condition 3.9 with $r_\mu = 0$; namely, $\lim_{t\to\infty} tS'(A_\mu(t)) = 0$ $(1 \le \mu \le m)$, and they satisfy the Growth Condition 3.3 with $a \le 0$; namely,

$$\lim_{t\to\infty}\,[at + A_\mu(t)] = \begin{cases} +\infty \text{ if } a = 0 \\ \\ -\infty \text{ if } a < 0 \end{cases} \quad .$$

For some sigmoids there are several choices of functions that satisfy the above two limits with different rates of convergence, and in some applications it may be advantageous to select different functions A_μ in order to satisfy Inequalities 4.2, 4.3, 4.5, and 4.6 with smaller values of w_μ $(1 \le \mu \le m)$. The last two lemmas and the theorem were stated with sufficient generality to accommodate different functions A_μ. If, however, the functions A_μ $(1 \le \mu \le m)$ are all the same function, then the conditions of the lemmas can be simplified. Before closing this section, we present these simplifications and briefly discuss qualitatively when and why one would choose functions A_μ with different rates of convergence in the two limits above.

Proposition 4.1. Let S be an odd differentiable function such that on $(0, \infty)$, S' is nonincreasing and positive. Assume that $A_\mu = A$ for all $\mu = 1, 2, ..., m$. Let

$$a \equiv \min_{1 \le \mu \le m} v(x_\mu - x_{\mu-1}) \ .$$

If $w > 0$ satisfies the following three inequalities

$$A(w) > \alpha \tag{4.10}$$

$$-\frac{a}{2} w + A(w) < 0 \tag{4.11}$$

$$0 < w\, S'(A(w)) < \delta_2 \tag{4.12}$$

and $w_\mu = w$ for all $\mu = 1, 2, ..., m$, then Inequalities 4.2, 4.3, 4.5, and 4.6 hold for all $\mu = 1, 2, ..., m$. Here α and δ_2 are as in Lemmas 4.1 and 4.2, respectively.

Proof. Since $A_\mu = A$ and $w_\mu = w$ $(1 \le \mu \le m)$, clearly Inequality 4.10 implies Inequality 4.2 and Inequality 4.12 implies Inequality 4.5. Inequalities 4.10 and 4.11 together imply $-aw + A(w) < -A(w) < -\alpha$, which implies Inequality 4.3 for $1 \le \mu \le m$ by definition of a. Now, the next series of inequalities follows from the definition of a, $S' > 0$ and nondecreasing on $(-\infty, 0)$, Inequality 4.11, S' even, and Inequality 4.12:

$$\begin{aligned} 0 &< w_\mu S'(w_\mu v(x_{\mu-1} - x_\mu) + A_\mu(w_\mu)) \le w\, S'(-wa + A(w)) \le w\, S'(-A(w)) \\ &= w\, S'(A(w)) < \delta_2 \ . \end{aligned}$$

Therefore, Inequality 4.6 holds for all $\mu = 1, 2, ..., m$. ////

In Proposition 4.1 we are simply taking advantage of the fact that once Inequality 4.3 is satisfied for that μ that gives the smallest value of $v(x_\mu - x_{\mu-1})$, then the same value w_μ satisfies Inequality 4.3 for all the other values of μ. However, a very small value of $v(x_\mu - x_{\mu-1})$ may require an extremely large value of w_μ to satisfy Inequality 4.3, while the same inequality may be satisfied by more conservative values of w_μ for the other values of μ.

In cases where a large discrepancy exists between the terms $v(x_\mu - x_{\mu-1})$ $(1 \le \mu \le m)$, it may be better to satisfy each of the Inequalities 4.3 with different values for w_μ. Moreover, since the terms $w_\mu v(x_{\mu-1} - x_\mu)$ and $A_\mu(w_\mu)$ in Inequality 4.3 are competing against each other in the sense that $A_\mu(w_\mu)$ is **increasing** with w_μ while $w_\mu v(x_{\mu-1} - x_\mu)$ is **decreasing** linearly with w_μ, it may be advantageous to choose functions A_μ with different rates of divergence depending on the sizes of the terms $v(x_{\mu-1} - x_\mu)$ $(1 \le \mu \le m)$. Qualitatively speaking, a slow rate of divergence for A_μ implies a smaller value for w_μ in Inequality 4.3, but it also implies a larger value for w_μ in Inequality 4.2. A faster rate of divergence for A_μ, of course, would imply the opposite.

Finally, we should point out that the sizes of the terms $v(x_{\mu-1} - x_\mu)$ $(1 \le \mu \le m)$ also depend on the choice of v. How to choose v^T and the functions A_μ optimally will not be discussed here. These are issues that require further research. We do believe, however, that, as a general rule, the **faster** S converges to 1, the **slower** A_μ will grow and the smaller the weights will be.

The following simple examples illustrate some of the points mentioned above. Hopefully, they also will help the reader appreciate the simplicity of the technique for determining weights that follows from Theorem 4.1.

Throughout these examples the sigmoid will be the hyperbolic tangent: $S(t) = \tanh(t)$ with derivative $S'(t) = \mathrm{sech}^2(t)$ $(t \in \mathbf{R})$. It is not hard to show that, for any $\eta > 0$, the function

$$A_\eta(t) \equiv \cosh^{-1}\left(\sqrt{t^{1+\eta}}\right) \equiv \ln\left[\sqrt{t^{1+\eta}} + \sqrt{t^{1+\eta} - 1}\right] \qquad (t \ge 1)$$

satisfies

$$t\, S'(A_\eta(t)) = \frac{1}{t^\eta}\ , \quad \text{for all } t \ge 1.$$

The set $X \subset \mathbf{R}^m$ is $X = \prod_{i=1}^{m} (1, \infty)$. The functions A_η and S are strictly increasing. The algorithm that we shall use is based on Proposition 4.1. For each $a > 0$ and $\eta > 0$, let $g_{a,\eta}$ denote the function appearing on the left-hand side of Inequality 4.11; that is

$$g_{a,\eta}(t) = -\frac{1}{2}\, a\, t + A_\eta(t) \qquad (t \ge 1)\ .$$

Let $t^*_\eta(a)$ denote the value of t where $g_{a,\eta}$ achieves its maximum value. When $\eta = 1$ or $\eta = 3$, one can find a closed-form expression for $t^*_\eta(a)$:

$$t^*_1(a) = \sqrt{1 + \frac{4}{a^2}}\ , \quad t^*_3(a) = \sqrt{\frac{8}{a^2} + \sqrt{\left(\frac{8}{a^2}\right)^2 + 1}}\ , \quad (a > 0)\ . \qquad (4.13)$$

Since $g_{a,\eta}$ is decreasing on $[t_\eta^*(a), \infty)$, we shall look in this interval for a value of w that satisfies Inequality 4.11. Note that since $wS'(A_\eta(w)) = \frac{1}{w^\eta}$, Inequality 4.12 is satisfied by all $w > (1/\delta_2)^{1/\eta}$. Similarly, since A_η is an increasing function, Inequality 4.10 is satisfied by all $w > A_\eta^{-1}(\alpha)$, where A_η^{-1} denotes the inverse of the function A_η. Thus, the strategy will be to find $w \geq \max\{(1/\delta_2)^{1/\eta}, A_\eta^{-1}(\alpha), t_\eta^*(a)\}$ that satisfies Inequality 4.11. The inverse of A_η is given by

$$A_\eta^{-1}(\alpha) = [\cosh(\alpha)]^{\frac{2}{1+\eta}} \qquad (\alpha \geq 0) \quad .$$

Note that when a is small, one can use the following approximations:

$$t_1^*(a) \cong \frac{2}{a} \quad \text{and } t_3^*(a) \cong \frac{4}{a}, \qquad \text{for "small a"} \quad (0 < a < 1).$$

The Problem. Given $x_0, x_1, \ldots, x_m$ in $\mathbf{R}^n$, $y_0, y_1, \ldots, y_m$ in $\mathbf{R}^\ell$, $\varepsilon_1 > 0$, and $\varepsilon_2 > 0$, find $\boldsymbol{w} \in X$ such that

$$|\, [T_w(x_j) - y_j]_i \,| < \varepsilon_1 \qquad (1 \leq i \leq \ell,\ 0 \leq j \leq m)$$

and

$$|\, [T'_w(x_k)]_{ij} \,| < \varepsilon_2 \qquad (1 \leq i \leq \ell,\ 1 \leq j \leq n,\ 0 \leq k \leq m)$$

for all $w > \boldsymbol{w}$.

Algorithm 4.1.

Step 1.

1.1. Compute $M_1 \equiv \max\limits_{1 \leq j \leq \ell} \sum\limits_{i=0}^{m} |y_{ij}|$, ($y_{ij}$ = j^{th} component of y_i).

1.2. Choose $\delta_1 \in (0, 1)$ such that $\delta_1 < \frac{2}{3}\varepsilon_1 M_1^{-1}$.

1.3. Set $\alpha = S^{-1}(1 - \delta_1)$. ($S^{-1}$ denotes the inverse of S.)

Step 2.

2.1. Choose v^T in $\mathbf{R}^n$ at random.

2.2. Order the numbers vx_k $(0 \leq k \leq m)$ and relabel:

$$vx_{k_0} < vx_{k_1} < \ldots < vx_{k_m} \quad .$$

2.3. Compute the consecutive differences of the above numbers:

$a_\mu \equiv v(x_{k_\mu} - x_{k_{\mu-1}})$ $\quad (1 \leq \mu \leq m)$ and order the numbers a_μ:

$$0 < a_{\mu_1} \leq a_{\mu_2} \leq \ldots \leq a_{\mu_m} \quad .$$

If $a_{\mu_1} = 0$, repeat Step 2.

Step 3. (For Derivative Control)

3.1. Compute

$$M_2 \equiv \max_{\substack{1 \leq i \leq \ell \\ 1 \leq j \leq n}} \left| v_j \right| \sum_{\mu=1}^{m} \left| (y_{k_\mu} - y_{k_{\mu-1}})_i \right| \; .$$

3.2. Choose $\delta_2 > 0$ such that $\delta_2 < 2\,\varepsilon_2\, M_2^{-1}$.

Set $i = 1$.

Step 4.

4.1. Set $a = a_{\mu_i}$.

4.2. Choose $\eta > 0$ for appropriate decay rate of $tS'(A_\eta(t)) \equiv \frac{1}{t^\eta}$ · (Note: Small a calls for low rate of increase of A_η, thus low rate of decay of $\frac{1}{t^\eta}$; i.e., small η.)

4.3. Set $A = A_\eta$.

4.4. Let $t_\alpha \equiv A^{-1}(\alpha)$, $t_{\delta_2} = (\frac{1}{\delta_2})^{1/\eta}$; if t_{δ_2} is too large, increase η.

4.5. Let t^* = value of t where $[-\frac{a}{2}\, t + A(t)]$ attains its maximum value.

4.6. Set $t = \max\,\{t_\alpha, t_{\delta_2}, t^*\}$.

(Note: The value $\boldsymbol{t}$ has the property that Inequalities 4.10 and 4.12 hold for all $w > \boldsymbol{t}$ and the function $[-\frac{a}{2}\, t + A(t)]$ is decreasing for $t > \boldsymbol{t}$.)

4.7. If $-\frac{1}{2} a\boldsymbol{t} + A(\boldsymbol{t}) \geq 0$, let $t = (2A(\boldsymbol{t}) + q)/a$, let $\boldsymbol{t} = t$, and repeat Step 4.6. (Note: Here $q > 0$ and $q \cong 1$. The larger q is, the faster the convergence to a value $\boldsymbol{t}$ satisfying $-\frac{1}{2} a\boldsymbol{t} + A(\boldsymbol{t}) < 0$; however, too large a q can lead to an excessively large $\boldsymbol{t}$.)

4.8. If $-\frac{1}{2} a\boldsymbol{t} + A(\boldsymbol{t}) < 0$, set $w_{\mu_i} = \boldsymbol{t}$.

4.9. If $i = m$, stop.

4.10. If w_{μ_i} is not too large, set $w_{\mu_j} = w_{\mu_i}$ for all $j \geq i$. Stop.

4.11. Set $i = i + 1$. Repeat Step 4.

Remark 4.2. Since the function $g(t) \equiv -\frac{1}{2} a\boldsymbol{t} + A(\boldsymbol{t})$ is decreasing on the interval (t^*, ∞) and $\lim_{t\to\infty} g(t) = -\infty$ when $a > 0$, it is easy to see that the iteration in Step 4.6 will yield a value $\boldsymbol{t}$ such that $g(\boldsymbol{t}) < 0$ in a finite number of steps whenever $q > 0$. To see this, assume $g(t^*) \geq 0$ and let $T \geq t^*$ satisfy $g(T) = 0$. Set $t_{k+1} = (2A(t_k) + q)/a$ $(k = 0, 1, 2, \ldots)$, where t_0 is any point in $[t^*, T]$. Now, if $t_k \in [t^*, T]$, then $g(t_k) \geq 0$; thus, $t_{k+1} \geq t_k + \frac{q}{a}$. Therefore, as long as t_0 and $t_k \in [t^*, T]$, we have $t_k \geq t_0 + k\frac{q}{a}$. This means that t_k cannot be less than T for all $k > 0$. Thus, after a finite number of iterations, t_k leaves the interval $[t^*, T]$ and $g(t_k) < 0$. ////

In the following examples, η will be either 1 or 3 so that we may determine $t^*_\eta(a)$ from Equations 4.13.

Example 4.1a. The points of interpolation are {(0, 0), (1, 1), (1.1, -1), (2, 0)}. So let $x_0 = 0$, $x_1 = 1$, $x_2 = 1.1$, $x_3 = 2$, and $y_0 = 0$, $y_1 = 1$, $y_2 = -1$, $y_3 = 0$. Let $\varepsilon_1 = \varepsilon_2 = 0.001$.

Step 1.

$M_1 = 2, \quad \delta_1 < \frac{\varepsilon_1}{3} = 0.00033$
$\delta_1 = 0.0003$
$\alpha = S^{-1}(1 - \delta_1) = 4.402$

Step 2.

Since $x_0 < x_1 < x_2 < x_3$, let $v = 1$, so $k_\mu = \mu \quad (1 \le \mu \le 3)$
$a_1 = x_1 - x_0 = 1$
$a_2 = x_2 - x_1 = 0.1$
$a_3 = x_3 - x_2 = 0.9$
Since $a_2 < a_3 < a_1$, we have $\mu_1 = 2, \quad \mu_2 = 3, \quad \mu_3 = 1$.

Step 3.

$M_2 = |y_1 - y_0| + |y_2 - y_1| + |y_3 - y_2| = 1 + 2 + 1 = 4$
$\delta_2 < 2\varepsilon_2/4 = 0.0005$
$\delta_2 = 0.0004$
Let $i = 1$.

Step 4.

$a = a_{\mu_1} = a_2 = 0.1$
Since a is "small," choose "small" η.
Let $\eta = 1$
$A(t) = A_1(t) = \cosh^{-1}(t) \quad (t \ge 1)$
$A^{-1}(\alpha) = \cosh(\alpha), \quad (\alpha \ge 0)$
$t_\alpha = \cosh(\alpha) = \cosh(4.402) = 40.81$
$t_{\delta_2} = 1/\delta_2 = 2500$. This value of t_{δ_2} is excessively large. We must increase η.
Let $\eta = 3$
$A(t) = A_3(t) = \cosh^{-1}(t^2)$
$A^{-1}(\alpha) = \sqrt{\cosh(\alpha)}$
$t_\alpha = \sqrt{\cosh(\alpha)} = 6.38$ [smaller value of t_α means that A(t) is increasing faster]
$t_{\delta_2} = (1/\delta_2)^{1/3} = \sqrt[3]{2500} = 13.57$. This value of t_{δ_2} is acceptable.
$t^* = t_3^*(a) \cong 4/a = 40$
$t = 40$.

The following table shows the results of the iterations involved in Step 4.6.
We shall use $q = 1$ and $g(t) = -\frac{a}{2}t + A(t)$.

t	$-\frac{a}{2}t$	$A(t)$	$g(t)$	$t_{new} = 20\ A(t) + 10$
40	-2	8.07	6.07	171.4
171.4	-8.57	10.98	2.41	229.6
229.6	-11.48	11.57	-0.09	////

Let $w_{\mu_1} = w_2 = 230$.

Since a_2 is smaller than all the other a_{μ_j}, the value of w_2 will work for all of the other weights; however, we consider w_2 too large, so we will repeat Step 4 with $i = 2$.

Step 4 with i = 2.

$a = a_{\mu_2} = a_3 = 0.9$

Set $\eta = 3$ (in order to meet the derivative requirement with an acceptable value of w_3)

t_α and t_{δ_2} are as before (because η did not change)

$t^* = t_3^*(a) \approx \frac{4}{a} = 4.45$

$t = 13.57$ and $g(t) < 0$.

Let $w_{\mu_2} = w_3 = 13.57$.

Note that since $t = t_{\delta_2}$, it is the derivative requirement that will determine all of the remaining weights (i.e., w_1) even if the remaining a_{μ_i} are much larger than a_3.

Let $w_1 = w_3 = 13.57$. Stop.

The vector $w \approx [13.57, 230, 13.57]$ satisfies Theorem 4.1 for the data of this example.

To complete the example, we shall determine a neural net mapping $T_w : \mathbf{R} \to \mathbf{R}$ that interpolates through the data with an error less than $\varepsilon_1 = 0.001$ and with derivative less than $\varepsilon_2 = 0.001$ at the interpolation points. We shall use $w = w$.

Recall that $T_w(x) = y_0 + \frac{1}{2} Y\,[S_m(L_w(x)) - S_m(L_w(x_0))]$, where

$$L_w(x) = \begin{bmatrix} w_1(x - x_1) + A(w_1) \\ w_2(x - x_2) + A(w_2) \\ w_3(x - x_3) + A(w_3) \end{bmatrix}, \quad Y = [y_1 - y_0 \ \vdots \ y_2 - y_1 \ \vdots \ y_3 - y_2] .$$

Note that the function A in the i^{th} component of $L_w(x)$ is the function used in the computation of w_i $(1 \leq i \leq 3)$. The result is

$$T_w(x) = \frac{1}{2}[S(13.57x - 7.67) - 2S(230x - 241.43) + S(13.57x - 21.24)] \quad .$$

////

Example 4.1b. This example is the same as Example 4.1a. We want to show that, by working directly with Inequalities 4.3 and 4.6 instead of the shortcut presented in Proposition 4.1, Theorem 4.1 may be satisfied with smaller weights. We focus on the second weight $\boldsymbol{w}_2$. A simple calculation shows that w_2 = 185 satisfies Inequalities 4.2, 4.3, 4.5, and 4.6 with $A_\mu(t) = \cosh^{-1}(t^2)$, i.e., $\eta = 3$. Moreover, $\boldsymbol{w}$ = [13.57, 185, 13.57] satisfies Theorem 4.1. Note that $\boldsymbol{w}_2$ is smaller than in Example 4.1a. ////

Example 4.1c. Now let us consider Example 4.1a without the requirement on the derivative. Recall that in Step 4 we were forced to increase η from 1 to 3 in order to satisfy the requirement on the derivative. Without this requirement, we can use η = 1 to solve the interpolation problem with a smaller weight w_2. Again, we only focus on the second weight and the iterations involved in Step 4.6 with q = 1 and

$A(t) = A_1(t) = \cosh^{-1}(t)$
$\alpha = 4.402$
$t_\alpha = 40.81$
$t_1^*(a) \cong 2/a = 20$
$\boldsymbol{t} = 41.$

t	$-t/20$	$A(t)$	$g(t)$	$t_{new} = 20\,A(t) + 10$
41	-2.05	4.406	2.36	98.12
98.12	-4.91	5.279	0.37	115.6
115.6	-5.78	5.44	-0.33	////

The interpolation problem can be solved with w_2 = 116. Moreover, if we work directly with Inequality 4.3, we find that w_2 = 100 also will solve the interpolation problem. ////

Example 4.2. The six inputs of this example belong to $\mathbf{R}^3$; x_0 through x_5 are, respectively, $[0\ \ 0\ \ 0]^T$, $[0\ \ 1\ \ 0]^T$, $[1\ \ 0\ \ 0]^T$, $[1\ \ 1\ \ 0]^T$, $[0\ \ 0\ \ 0.1]^T$, and

$[0\ 0\ 1]^T$. The outputs y_0 through y_5 belong to $\mathbf{R}^2$; they are $[0\ \ 0]^T$, $[1\ \ 0]^T$, $[0\ \ 1]^T$, $[0\ \ 0]^T$, $[1\ \ 1]^T$, and $[0\ \ 3]^T$. We wish to interpolate through the points (x_i, y_i), $0 \leq i \leq 5$ with an error bound $\varepsilon_1 = 0.001$ and a derivative less than $\varepsilon_2 = 0.01$ at the interpolation points.

Step 1.

$M_1 = \max\{2, 5\} = 5, \quad \delta_1 < \frac{2\varepsilon_1}{15} = 0.00013$
$\delta_1 = 0.0001$
$\alpha = S^{-1}(1 - \delta_1) = 4.952.$

Step 2.

Let $v = [2\ 1\ {-1}]$ to get
$vx_0 = 0$, $vx_1 = 1$, $vx_2 = 2$, $vx_3 = 3$, $vx_4 = -0.1$, $vx_5 = -1$.
Since $vx_5 < vx_4 < vx_0 < vx_1 < vx_2 < vx_3$, we have
$k_0 = 5$, $k_1 = 4$, $k_2 = 0$, $k_3 = 1$, $k_4 = 2$, $k_5 = 3$.
$a_1 = v(x_{k_1} - x_{k_0}) = vx_4 - vx_5 = -0.1 - (-1) = 0.9$
$a_2 = v(x_{k_2} - x_{k_1}) = vx_0 - vx_4 = 0 - (-0.1) = 0.1$
$a_3 = v(x_{k_3} - x_{k_2}) = vx_1 - vx_0 = 1 - 0 = 1$
$a_4 = v(x_{k_4} - x_{k_3}) = vx_2 - vx_1 = 2 - 1 = 1$
$a_5 = v(x_{k_5} - x_{k_4}) = vx_3 - vx_2 = 3 - 2 = 1$
Since $a_2 < a_1 < a_3 = a_4 = a_5$ we have
$\mu_1 = 2$, $\mu_2 = 1$, $\mu_3 = 3$, $\mu_4 = 4$, $\mu_5 = 5$.

Step 3.

Consider the matrix $M \equiv \sum_{\mu=1}^{4} | y_{k_\mu} - y_{k_{\mu-1}} | \, | v |$.

$$M \equiv \Big[| y_4 - y_5 | + | y_0 - y_4 | + | y_1 - y_0 | + | y_2 - y_1 | + | y_3 - y_2 | \Big] \Big[2\ 1\ 1 \Big]$$
$$= \begin{bmatrix} 4 \\ 5 \end{bmatrix} [2\ 1\ 1] = \begin{bmatrix} 8 & 4 & 4 \\ 10 & 5 & 5 \end{bmatrix}$$

By inspection, we get $M_2 = 10$, so $\delta_2 < \frac{2\,\varepsilon_2}{10} = 0.002.$

Let $\delta_2 = 0.001$
Let $i = 1$.

Step 4.

$a = a_{\mu_1} = a_2 = 0.1$

Let $\eta = 3$, $A(t) = A_3(t) = \cosh^{-1}(t^2)$, $A^{-1}(\alpha) = \sqrt{\cosh(\alpha)}$
$t_\alpha = 8.41$
$t_{\delta_2} = (1/\delta_2)^{1/3} = 10$
$t^* = t_3^*(a) = 40$

$w_2 = 230$ (see Example 4.1a.)
Let i = 2.

Step 4.

$a = a_{\mu_2} = a_1 = 0.9$
$\eta = 3$
$t_\alpha = 8.41$, $t_{\delta_2} = 10$ (as before)
$t^* = t_3^*(a) \cong \dfrac{4}{a} = 4.45$
$t = \max\{t_\alpha, t_{\delta_2}, t^*\} = t_{\delta_2} = 10$

t	$-0.45t$	$A(t)$	$g(t)$	$t_{new} = 2.22\ A(t) + 1.11$
10	-4.5	5.29	0.8	12.85
12.85	-5.78	5.8	0.015	13.98
13.98	-6.29	5.97	-0.32	////

$w_1 = 14$
Let i = 3.

Step 4.

$a = a_{\mu_3} = a_3 = 1$
$\eta = 3$
$t_\alpha = 8.41$ and $t_{\delta_2} = 10$ (as before)
$t^* = t_3^*(a) \cong \dfrac{4}{a} = 4$
$t = 10$

t	$-0.5t$	$A(t)$	$g(t)$	$t_{new} = 2\,A(t) + 1$
10	-5.0	5.29	0.29	11.58
11.58	-5.79	5.59	-0.2	////

Set w_3 = 11.6 and all the rest of the weights = 11.6
w_4 = 11.6
w_5 = 11.6
w = $[14,\ 230,\ 11.6,\ 11.6,\ 11.6]^T$. ////

The last example illustrates the modular features of the technique and shows how to estimate the error directly from Lemma 4.1.

Example 4.3. Consider the "exclusive or" problem; that is, a map that interpolates through (x_i, y_i), i = 0, 1, 2, 3, where $x_0 = [0 \ \ 0]^T$, $x_1 = [0 \ \ 1]^T$, $x_2 = [1 \ \ 0]^T$, $x_3 = [1 \ \ 1]^T$, $y_0 = 0 = y_3$, and $y_1 = y_2 = 1$. We can assemble a network that implements the exclusive or problem using "part" of the network in Example 4.2.

Note that the matrix v = [2 1] maps the vector x_i into i for each i = 0, 1, 2, 3; note that the matrix v of the previous example achieved the same result. Thus, we already have a network (with three hidden units) in the previous network that maps the integers 0, 1, 2, 3 into desired outputs y_0, y_1, y_2, y_3. Consequently, all we need to do is to use the correct matrix Y. If v = [2 1], $w = [11.6,\ 11.6,\ 11.6]^T$, and Y = [1 0 -1], we have a net that implements the exclusive or problem.

Let us use Lemma 4.1 to estimate the error. Since $w_i = 11.6$ and $A(w_i) = 5.59$, we know that Inequalities 4.2 and 4.3 hold with $\alpha = 5.58$. Since $1 - S(\alpha) = 0.000028$, we conclude that Inequality 4.4 holds with $\delta_1 = 0.00003$. By Lemma 4.1, the error is bounded by $\frac{3}{2}\delta_1 \sum_{i=0}^{3} | y_i | = 3\delta_1 = 0.00009$. The mapping is given by

$$T(x) = \frac{1}{2}\,[S(11.6\ vx - 6.01) - S(11.6\ vx - 29.21)] - 0.000005 \qquad (x \in \mathbf{R}^2),$$

with v = [2 1].

$$T(x_0) = 0, \quad T(x_1) = 0.99998 = T(x_2), \quad T(x_3) = 0.000009. \qquad ////$$

Remark 4.3. The reader might have noticed that when the value of η is determined and fixed by the requirements on the derivative at the interpolation points, as it was the case in some of the examples above, then the algorithm is more efficient if Steps 4.2 through 4.4 are performed immediately after Step 3.2, for then those steps are performed only once. ////

REFERENCES

1. E. D. Sontag. "Capabilities and Training of Feedforward Nets," in *Theory and Applications of Neural Networks*, R. Mammone and J. Zeevi, eds. New York, Academic Press, 1991.

2. M. Arai. "Mapping Abilities of Three-Layer Neural Networks," in *Proceedings of the International Joint Conference on Neural Networks*. Washington, D.C., IEEE Publications, 18-22 June 1989. Pp. I-419-24.

3. D. Chester. "Why Two Hidden Layers Are Better Than One," in *Proceedings of the International Joint Conference on Neural Networks*. Washington, D.C., IEEE Publications, January 1990. Pp. I-265-68.

4. Naval Weapons Center. *On the Interpolation Properties of Feedforward Layered Neural Networks*, by J. M. Martin. China Lake, Calif., NWC, October 1990. 31 pp. (NWC TP 7094, publication UNCLASSIFIED.)

5. H. Guo and S. B. Gelfand. "Analysis of Gradient Descent Learning Algorithms for Multilayer Feedforward Neural Networks," *IEEE Trans. Circuits Syst.*, Vol. 38, No. 8 (August 1991), pp. 883-94.

6. J. F. Kolen and A. K. Goel. "Learning in Parallel Distributed Processing Networks: Computational Complexity and Information Content," *IEEE Trans. Syst., Man, Cybern.*, Vol. 21, No. 2 (March/April 1991), pp. 359-68.

7. M. W. Hirsch and S. Smale. *Differential Equations, Dynamical Systems, and Linear Algebra*. New York, Academic Press, 1974.

8. W. Rudin. *Real and Complex Analysis*. New York, McGraw-Hill, 1974.

APPENDIX

Proof of Lemma 3.1. Let ℓ_{ij} denote the unique line through x_i and x_j and let H_{ij} be the hyperplane through the origin consisting of all vectors in $\mathbf{R}^n$ that are perpendicular to ℓ_{ij} for $0 \leq i \leq m$, $0 \leq j \leq m$, and $i \neq j$. Let $H \equiv \cup \{H_{ij} : i \neq j, 0 \leq i \leq m, 0 \leq j \leq m\}$. If $v^T \notin H$, then $\{vx_i : 0 \leq i \leq m\}$ is a set of distinct numbers; for if not, say $vx_i = vx_j$, then $v(x_i - x_j) = 0$, which implies that v^T is perpendicular to the line ℓ_{ij}; thus, $v^T \in H$, a contradiction. ////

Remark A.1. Since the set H has Lebesgue measure zero, it follows that all vectors v^T in $\mathbf{R}^n$ satisfy the hypothesis of the lemma except for those on a set of measure zero. ////

Proof of Lemma 3.2. We must show that for every $r_k \geq 0$

$$\lim_{t \to \infty} A_k(t) = \infty \tag{A-1}$$

and

$$\lim_{t \to \infty} t\, S'(at + A_k(t)) = 0 \qquad \text{for all } a \neq 0 \ , \tag{A-2}$$

where A_k satisfies Condition 3.9 and Condition 3.3 with $a < 0$.

Fix $r_k \geq 0$ and consider $A_k : (t_k, \infty) \to (0, \infty)$. If Equation A-1 does not hold, there exists $M > 0$ and an unbounded sequence $\mu_1 < \mu_2 < ... < \mu_n < ...$ such that $A_k(\mu_n) \leq M$ for all $n \geq 1$. Since S' is nonincreasing on $(0, \infty)$, we have $S'(A_k(\mu_n)) \geq S'(M)$ for all $n \geq 1$. Consequently,

$$\mu_n S'(A_k(\mu_n)) \geq \mu_n S'(M) \to \infty \text{ as } n \to \infty \ ,$$

which contradicts Condition 3.9. Therefore, Equation A-1 holds.

To prove Equation A-2, first assume that $r_k = 0$. Since S' is nonincreasing on $(0, \infty)$, $S'(at + A_k(t)) \leq S'(A_k(t))$ when $a > 0$ and $t > \max\{0, t_k\}$. Thus, when $a > 0$, Equation A-2 trivially follows from Condition 3.9 with $r_k = 0$. When $a < 0$, Condition 3.3 gives

$$\lim_{t \to \infty} \left[\frac{1}{2} at + A_k(t) \right] = -\infty \ .$$

Certainly, then, there exists $T \geq t_k$ such that $\frac{1}{2}$ at $+ A_k(t) < 0$ for all $t > T$; that is, $-ta - A_k(t) > A_k(t) > 0$ for all $t > T$. Since S' is an even function and nonincreasing on $(0, \infty)$, it follows from the last inequality that

$$tS'(ta + A_k(t)) = t\, S'(-ta - A_k(t)) \leq tS'(A_k(t)) \qquad \text{for all } t > T.$$

Hence, when $a < 0$, Equation A-2 also follows from Condition 3.9 with $r_k = 0$.

Next, fix $r_k > 0$. To establish Equation A-2, we shall show that to every $\varepsilon > 0$ there corresponds a $T \geq t_k$ such that $tS'(ta + A_k(t)) < \varepsilon$ for all $t > T$.

If $r_o \equiv \frac{\varepsilon}{2}$, the hypothesis of the lemma gives us a function $A_o : (t_o, \infty) \rightarrow (0, \infty)$ such that

$$t\, S'(A_o(t)) = \frac{\varepsilon}{2} \qquad \text{for all} \qquad t > t_o \ . \tag{A-3}$$

If $a > 0$, Condition 3.3 applied to A_o shows that there exists $T_o \geq t_o$ such that $t(-a) + A_o(t) < 0$ or, equivalently, $ta > A_o(t)$ for all $t > T_o$. Since S' is nonincreasing on $(0, \infty)$ and A_k is positive valued, it follows from the last inequality that

$$S'(ta + A_k(t)) \leq S'(A_o(t)) \qquad \text{for all} \qquad t > \max\{T_o, t_k\} \ .$$

Let $T \equiv \max\{T_o, t_k\}$. Equation A-3 and the last inequality imply $tS'(ta + A_k(t)) \leq \frac{\varepsilon}{2} < \varepsilon$ for all $t > T$. This proves Equation A-2 for $a > 0$ and $r_k > 0$.

Now assume that $a < 0$. Condition 3.3 applied to A_k and A_o shows that there exist $T_1 > t_k$ and $T_2 > t_o$ such that

$$\frac{1}{2}at + A_k(t) < 0 \qquad \text{for all} \qquad t > T_1$$

$$\frac{1}{2}at + A_o(t) < 0 \qquad \text{for all} \qquad t > T_2 \ .$$

Consequently, if $T \equiv \max\{T_1, T_2\}$, then $-ta - A_k(t) > A_o(t) > 0$ for all $t > T$. And, as before, since S' is an even function and nonincreasing on $(0, \infty)$, we have

$$tS'(at + A_k(t)) = tS'(-at - A_k(t)) \leq tS'(A_o(t)) = \frac{\varepsilon}{2} < \varepsilon \qquad \text{for } t > T.$$

Therefore, Equation A-2 holds when $a < 0$ and $r_k > 0$. This completes the proof of Lemma 3.2. ////

Remark A-2. For the purpose of this remark, denote by A_r a function $A_r : (t_r, \infty) \to (0, \infty)$ that satisfies

$$t\, S'(A_r(t)) \equiv r \qquad \text{for all} \qquad t > t_r \ , \tag{A-4}$$

where $r > 0$, and let $A_0 : (t_0, \infty) \to (0\ \infty)$ satisfy

$$\lim_{t \to \infty} t\, S'(A_0(t)) = 0 \ .$$

It is easy to see that the functions A_r above are nondecreasing for all $r > 0$ whenever S' is nonincreasing on $(0, \infty)$. This fact was not needed in the development of the theory in Section 3; however, it may prove useful when implementing the techniques presented in this paper. To see that A_r is nondecreasing when $r > 0$ assume otherwise; assume there exist $a < b$, both in (t_r, ∞), such that $A_r(a) > A_r(b)$. Then, $S'(A_r(a)) \leq S'(A_r(b))$, which implies $r = a\, S'(A_r(a)) \leq a\, S'(A_r(b)) < b\, S'(A_r(b)) = r$, a contradiction.

The function A_0 can be defined in such a way that it too is a nondecreasing function, provided t_r does not increase as r decreases. This can be done as follows. Suppose that, for $0 < r \leq r_0$, t_r does not increase as r decreases. Let $f : (t_{r_0}, \infty) \to (0, r_0)$ be a **decreasing** function such that $\lim_{t \to \infty} f(t) = 0$. Define $A_0 : (t_{r_0}, \infty) \to (0, \infty)$ by

$$A_0(t) \equiv A_{f(t)}(t) \qquad \text{for} \qquad t > t_{r_0} \ .$$

Note that $f(t) < r_0$ for all $t > t_{r_0}$ implies $t_{f(t)} \leq t_{r_0}$ for all $t > t_{r_0}$. Therefore, $A_{f(t)}(t)$ is well defined for all $t > t_{r_0}$; that is, t is in the domain of $A_{f(t)}$ for all $t > t_{r_0}$. We claim that A_0 is a nondecreasing function. The proof is by contradiction: if $A_0(a) > A_0(b)$ for some $a < b$ in the domain of A_0, then, by definition of A_0, we have $A_{f(a)}(a) > A_{f(b)}(b)$. Since S' is nonincreasing on $(0, \infty)$, with the aid of Equation A-4 we conclude

$$f(a) = a\, S'(A_{f(a)}(a)) \leq a\, S'(A_{f(b)}(b)) < b\, S'(A_{f(b)}(b)) = f(b) \ ,$$

which contradicts the fact that f is a decreasing function. Note that $\lim_{t\to\infty} tS'(A_o(t)) = \lim_{t\to\infty} tS'(A_{f(t)}(t)) = \lim_{t\to\infty} f(t) = 0.$ ////

Proof of Lemma 4.1. Fix $j \in \{0, 1, 2, ..., m\}$ and $w \in X$. Assume w satisfies Inequalities 4.2 and 4.3. Let $z_j \equiv S_m(L_w(x_j)) - S_m(L_w(x_o))$ and let z_{jk} denote the k^{th} component of z_j $(1 \le k \le m)$. Set $z_{jo} \equiv 2$ and $z_{j(m+1)} \equiv 0$. Equation 4.1 gives

$$T_w(x_j) - y_j = \left[y_o + \frac{1}{2} Y z_j\right] - y_j = y_o + \frac{1}{2}\sum_{k=1}^{m} z_{jk}(y_k - y_{k-1}) - y_j$$

$$= \begin{cases} \dfrac{1}{2}[\sum\limits_{\substack{k=0\\k\neq j}}^{m} y_k(z_{jk} - z_{j(k+1)}) + y_j(z_{jj} - z_{j(j+1)} - 2)] & \text{if } j < m \\ \\ \dfrac{1}{2}\left[\sum\limits_{k=o}^{m-1} y_k(z_{jk} - z_{j(k+1)}) + y_m(z_{mm} - 2)\right] & \text{if } j = m \ . \end{cases} \quad \text{(A-5)}$$

Hence, it suffices to prove Inequalities A-6 through A-9 for $1 \le j \le m$. Note that the interpolation through (x_o, y_o) is exact.

$$|2 - z_{j1}| < 2\delta_1 \quad \text{(A-6)}$$

$$|z_{jk} - z_{j(k+1)}| < 2\delta_1 \qquad \text{for } 1 \le k < m, k \neq j \quad \text{(A-7)}$$

$$|z_{jj} - z_{j(j+1)} - 2| < 3\delta_1 \qquad \text{if } j < m \quad \text{(A-8)}$$

$$|z_{jm}| < \delta_1 \qquad \text{if } j < m \text{ , and } \quad |z_{mm} - 2| < 2\delta_1 \ . \quad \text{(A-9)}$$

Clearly, Inequalities A-5 through A-9 imply Inequality 4.4.

Since $v(x_j - x_k) \ge 0$ for $1 \le k \le j$, Inequality 4.2 implies

$$w_k v(x_j - x_k) + A_k(w_k) > \alpha \qquad \text{for } 1 \le k \le j \ .$$

Therefore, by the choice of α, the k^{th} component of $S_m(L_w(x_j))$ is within δ_1 of 1 for $1 \leq k \leq j$; that is,

$$1 - \delta_1 < S(w_k v(x_j - x_k) + A_k(w_k)) < 1 \qquad \text{for } 1 \leq k \leq j \ . \qquad \text{(A-10)}$$

Since $v(x_j - x_k) \leq v(x_{k-1} - x_k)$ for $0 \leq j < k \leq m$, Inequality 4.3 implies

$$w_k v(x_j - x_k) + A_k(w_k) < -\alpha \qquad \text{for } 0 \leq j < k \leq m \ .$$

Therefore, by the choice of α, the k^{th} component of $S_m(L_w(x_j))$ is within δ_1 of -1 for $0 \leq j < k \leq m$; that is,

$$-1 < S(w_k v(x_j - x_k) + A_k(w_k)) < -1 + \delta_1 \qquad \text{for } 0 \leq j < k \leq m \ . \qquad \text{(A-11)}$$

The definitions of β, L_w, and z_j give

$$\begin{aligned} |z_{jk} - z_{j(k+1)}| \leq {} & |S_m(L_w(x_j))_k - S_m(L_w(x_j))_{k+1}| + \\ & |S_m(L_w(x_o))_k - S_m(L_w(x_o))_{k+1}| \qquad \text{for } 1 \leq k < m \ , \end{aligned} \qquad \text{(A-12)}$$

where $S_m(L_w(x_j))_k$ denotes the k^{th} component of $S_m(L_w(x_j))$ for all k and j.

The second term on the right-hand side (RHS) of Inequality A-12 is less than δ_1 for $1 \leq k < m$ as Inequality A-11 with $j = 0$ shows. The first term on the RHS of Inequality A-12 is less than δ_1, as shown by Inequality A-10 when $1 \leq k < j$ and by Inequality A-11 when $j < k < m$. This proves Inequality A-7.

Next, note that

$$\begin{aligned} |z_{jj} - z_{j(j+1)} - 2| \leq {} & |S_m(L_w(x_j))_j - S_m(L_w(x_j))_{j+1} - 2| + \\ & |S_m(L_w(x_o))_j - S_m(L_w(x_o))_{j+1}| \ . \end{aligned} \qquad \text{(A-13)}$$

As before, the second term on the RHS of Inequality A-13 is less than δ_1 if $j < m$. By Inequality A-10 with $k = j$ and Inequality A-11 with $k = j + 1$, one concludes that

$$2 - 2\delta_1 < S_m(L_m(x_j))_j - S_m(L_w(x_j))_{j+1} < 2 \ ,$$

which shows that the first term on the RHS of Inequality A-13 is less than $2\delta_1$. This proves Inequality A-8.

Consider now z_{j1}:

$$0 < 2 - z_{j1} = 2 - [S_m(L_w(x_j))_1 - S_m(L_w(x_o))_1] < 2 - (1 - \delta_1) + (-1 + \delta_1) ,$$

where we used Inequality A-10 with k = 1 and Inequality A-11 with j = 0. This proves Inequality A-6.

If j < m, then Inequality A-11 implies

$$|z_{jm}| = |S_m(L_w(x_j))_m - S_m(L_w(x_o))_m| < \delta_1 .$$

If j = m, then Inequality A-10 with k = j = m and Inequality A-11 with j = 0 and k = m give

$$0 > z_{mm} - 2 = S_m (L_w(x_m))_m - S_m(L_w(x_o))_m - 2 > (1 - \delta_1) + (1 - \delta_1) - 2 = - 2\delta_1 .$$

Therefore Inequality A-9 holds. This completes the proof of Lemma 4.1. ////

Proof of Lemma 4.2. If $T_{w}(x)$ is given by Equation 4.1 then, by Equation 2.1,

$$[T'_w (x_k)]_{ij} = \frac{1}{2} Y_i S'_m(L_w(x_k))wv_j \qquad (0 \leq k \leq m, 1 \leq i \leq \ell, 1 \leq j \leq n) ,$$

where Y_i denotes the i^{th} row of the matrix Y $(1 \leq i \leq \ell)$ and v_j denotes the j^{th} component of v $(1 \leq j \leq n)$. This leads to

$$[T'_w(x_k)]_{ij} = \frac{1}{2} \sum_{\mu=1}^{m} (y_\mu - y_{\mu-1})_i w_\mu S'(w_\mu v(x_k - x_\mu) + A_\mu(w_\mu)) v_j .$$

Hence, if suffices to show

$$0 < w_\mu S'(w_\mu v(x_k - x_\mu) + A_\mu(w_\mu)) < \delta_2 \qquad (1 \leq \mu \leq m, 0 \leq k \leq m) . \text{(A-14)}$$

Inequality A-14 reduces to Inequality 4.5 when $k = \mu$ $(1 \leq \mu \leq m)$, and it is implied by Inequality 4.5 when $k > \mu$ $(1 \leq \mu < m)$ because S' is nonincreasing on $(0, \infty)$. Finally, since

$$w_\mu v(x_k - x_\mu) + A_\mu(w_\mu) \leq w_\mu v(x_{\mu-1} - x_\mu) + A_\mu(w_\mu) \qquad \text{when } k < \mu$$

and S' is nondecreasing on $(-\infty, 0)$, Inequality 4.6 implies Inequality A-14 for $k < \mu$ $(1 \leq \mu \leq m)$. Note that we used the fact that S is an odd function (which implies S' is even). ////

INITIAL DISTRIBUTION

1 Commander in Chief, U. S. Pacific Fleet, Pearl Harbor (Code 325)
1 Commander, Third Fleet
1 Commander, Seventh Fleet
2 Naval Academy, Annapolis (Director of Research)
1 Naval War College, Newport
1 Air Force Intelligence Agency, Bolling Air Force Base (AFIA/INT, MAJ R. Esaw)
2 Defense Technical Information Center, Alexandria
1 Hudson Institute/Washington Office, Washington, DC (Technical Library)