On the Interpolation Properties of Feedforward Layered Neural Networks

by

Jorge M. Martin
Research Department

OCTOBER 1990

NAVAL WEAPONS CENTER
CHINA LAKE, CA 93555-6001

Approved for public release; distribution unlimited
FOREWORD

This report documents a collection of results concerning the interpolation properties of feed forward multilayered neural networks that were discovered in a research project jointly sponsored by the Office of Naval Research and the Independent Research Program of the Naval Weapons Center. The work was done at the Naval Weapons Center from December 1989 to June 1990 under Program Element 0601153N, Research Project RR014-05-0K, RR052-02-02, R&T Project Code 411p002...02, 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
24 Sep 90

Released for publication by
W. B. PORTER
Technical Director

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

NWC Technical Publication 7094

Published by ...................................................... Technical Information Department
Collation ................................................................. Cover, 17 leaves
First printing ............................................................ 86 Copies
**Title:** On the Interpolation Properties of Feedforward Layered Neural Networks

**Author:** Jorge M. Martin

**Performing Organization:** Naval Weapons Center
China Lake, CA 93555-6001

**Sponsoring Agency:** Office of Naval Research (Code 111)
800 N. Quincy Street
Arlington, VA 22217

**Abstract:**

The characterization of the interconnection weights of an L layered feedforward neural net that interpolates through a set of points is considered. A closed form expression for the last layer of weights of a net that interpolates through \( m_{L-1} + 1 \) points is derived in terms of the points of interpolation. These weights are a function of all the weights in the preceding layers which may be chosen at random, and \( m_{L-1} \) is the number of neurons in the layer preceding the output layer.

Another method for determining all the weights of a net with only two layers of weights is also presented. This method produces a transfer function that interpolates through \( m_0 + 1 \) points or less, where \( m_0 \) is the number of inputs to the net.

The norm of the Jacobian matrix of the transfer function at the interpolation points is introduced as a measure of the sensitivity of the transfer function to perturbations in the inputs of the interpolation points. The results suggest that small weights are required for low sensitivity.

**Subject Terms:**

Feedforward Layered Neural Network, interpolation, learning sensitivity, transfer function, Jacobian Matrix.
CONTENTS

1. Introduction ............................................................................................................. 3
2. Notation, Problem Statement, and Preliminary Results .................................. 4
3. The Last Layer of Weights: A Lower Bound for IC(\(\widetilde{m}\)) ..................... 9
4. Nets With Two Layers of Weights ......................................................................... 14
5. Comparisons and Applications ............................................................................. 16
6. Jacobian Matrix of the Transfer Function ............................................................ 18
7. Summary .................................................................................................................. 28
8. References ................................................................................................................ 31

ACKNOWLEDGMENT

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

Determining the interconnection weights of a feed-forward multilayered neural network so that the resulting transfer function (input-output map) will map a certain set of inputs to a corresponding set of desired outputs is viewed here as an interpolation problem. The layered net, which has \( L \geq 1 \) layers of weights, is described in the next section together with a statement of the interpolation problem and some preliminary results.

In Section 3 it is shown how one can interpolate through a set of \( m_{L-1} + 1 \) input-output points (or less) with distinct inputs, where \( m_{L-1} \) is the number of neurons in the layer preceding the output layer. This can be accomplished by a proper choice of the last layer of weights. A closed-form expression for these weights is given in terms of the \( m_{L-1} + 1 \) points of interpolation. These weights are a function of all of the weights in the preceding layers, which may be chosen at random.

Section 4 discusses nets with only two layers of weights \( (L = 2) \). A method is presented for determining all of its weights so that its transfer function interpolates through a set of \( m_0 + 1 \) points (or less), where \( m_0 \) is the number of neurons in the input layer. The two methods for selecting the weights are compared in Section 5, and suggestions for their applications are given.

The freedom that exists in the selection of the first \( L - 1 \) layers of weights when using the method of Section 3 can be used to reduce the sensitivity (to noisy input patterns) of the resulting input-output map. The sensitivity of the transfer function at an interpolation point is measured here by the norm of the Jacobian matrix (total derivative) of the transfer function at the given point. Since a small change in the input produces a change in the output whose magnitude is approximately bounded by the product of the norm of the
Jacobian matrix and the magnitude of the change in the input, it is suggested that by minimizing the norm of the Jacobian matrix at the interpolation points the change in output produced by a small change in the input can be minimized. Thus, an expression for the Jacobian matrix of the transfer function is derived and is presented in Section 6. Before computing its norm, the induced (p, q) matrix norms are introduced together with some of their properties. A judicious choice for p and q yields computable upper bounds for the norm of the Jacobian matrix. The results suggest that small weights are required for low sensitivity.

For an introduction to feed-forward layered neural nets (FLNNs) and some of their basic properties, the reader is referred to References 1 and 2.

2. NOTATION, PROBLEM STATEMENT, AND PRELIMINARY RESULTS

We will consider layered neural nets with architecture \( \hat{m} = (m_0, m_1, ..., m_L) \). This means (Reference 2) that the net consists of \( L + 1 \) layers of neurons with \( m_i \) neurons in the \( i^{th} \) layer \( (0 \leq i \leq L) \). The activation function of each neuron will be denoted by \( S: \mathbb{R} \rightarrow (-1, 1) \), where \( \mathbb{R} \) is the set of real numbers and \( S \) is assumed to be a strictly increasing continuous sigmoid mapping \( \mathbb{R} \) onto the open interval \((-1, 1)\). Thus, \( S \) is invertible with inverse \( S^{-1}: (-1,1) \rightarrow \mathbb{R} \). Euclidean n-space with the standard metric topology will be denoted by \( \mathbb{R}^n \), the set of real \( n \times m \) matrices will be denoted by \( \mathbb{R}^{n \times m} \), and \((-1, 1)^n = \{ x = [x_1, x_2, ..., x_n]^T \in \mathbb{R}^n \} \). For each positive integer \( n \), let \( S_n: \mathbb{R}^n \rightarrow (-1, 1)^n \) be defined by \( S_n(x) = [S(x_1), S(x_2), ..., S(x_n)]^T \), for all \( x = [x_1, x_2, ..., x_n]^T \in \mathbb{R}^n \). Similarly, let \( S_n^{-1}: (-1, 1)^n \rightarrow \mathbb{R}^n \) be defined by \( S_n^{-1}(x) = [S^{-1}(x_1), S^{-1}(x_2), ..., S^{-1}(x_n)]^T \), for all \( x = [x_1, x_2, ..., x_n]^T \in (-1, 1)^n \). The superscript \( T \) denotes transpose.

For each \( \beta \in \mathbb{R}^n \), let \( \hat{\beta}: \mathbb{R}^n \rightarrow \mathbb{R}^n \) denote the operator defined by \( \hat{\beta}(x) = x + \beta \), \( (x \in \mathbb{R}^n) \). If \( f \) and \( g \) are two functions with the range of \( g \) contained in the domain of \( f \), then \( f \circ g \) will denote the composition of \( f \) and \( g \). The same symbol
will be used to represent a linear transformation and its matrix with respect to
the standard orthonormal bases on its domain and range.

If $W_i \in \mathbb{R}^{m_i \times m_{i-1}} (1 \leq i \leq L)$ and $\beta_i \in \mathbb{R}^{m_i} (0 \leq i \leq L)$, let $T_i = (S_{m_i} \circ \hat{\beta}_i \circ W_i): \mathbb{R}^{m_{i-1}} \to (-1, 1)^{m_i} (1 \leq i \leq L)$ and let $T_0: \mathbb{R}^{m_0} \to \Lambda^{m_0}$ denote either the identity map [in which case, $\Lambda = \mathbb{R}$], the composition $S_{m_0} \circ \hat{\beta}_{m_0}$, or simply $S_{m_0}$ [in which cases, $\Lambda = (-1, 1)$]. Note that $T_0$ is injective; thus, one may define its inverse $T_0^{-1}$ as follows:

$$
T_0^{-1} = \begin{cases} 
T_0: \mathbb{R}^{m_0} \to \mathbb{R}^{m_0}, & \text{if } T_0 = \text{identity map} \\
S_{m_0}^{-1}: (-1, 1)^{m_0} \to \mathbb{R}^{m_0}, & \text{if } T_0 = S_{m_0} \\
(-\beta_0)^\wedge S_{m_0}^{-1}: (-1, 1)^{m_0} \to \mathbb{R}^{m_0}, & \text{if } T_0 = S_{m_0} \circ \hat{\beta}_0.
\end{cases}
$$

With this notation, the transfer function (input-output map) of the FLNN with architecture $\hat{m} = (m_0, m_1, ..., m_L)$ and activation function $S$ that will be considered here can be written as the composition

$$
\hat{F} = T_L \circ T_{L-1} \circ ... \circ T_1 \circ T_0: \mathbb{R}^{m_0} \to (-1, 1)^{m_L}.
$$

Remark 2.1. For the sake of generality, three possibilities have been allowed for $T_0$ in order to accommodate exceptions or variations in the interpretation or in the use of the first layer of neurons, which is considered by some authors as being only an "input layer" ($T_0 = \text{identity function}$), or may be used simply to "normalize the input" ($T_0 = S_{m_0}$), or to "normalize and center the input" ($T_0 = S_{m_0} \circ \hat{\beta}_0$). However, if $T_0 = S_{m_0} \circ \hat{\beta}_0$, it will be assumed that $\beta_0 \in \mathbb{R}^{m_0}$ is free to be chosen to satisfy some criterion other than the interpolation problem (e.g., to "center" the set of input data) and that once $\beta_0$ has been chosen, it remains fixed. Therefore, the only free parameters are $W_i$ and $\beta_i$, for $i = 1, 2, ..., L$.

Interpolation Problem 1 (IPI). Given a set of points of interpolation

$$
\Omega = \{(l_i, O_i) \in \mathbb{R}^{m_0 \times (-1, 1)^{m_L}}: 1 \leq i \leq k \text{ and } l_i \neq l_j \text{ for } i \neq j\},
$$

(2.2)
determine \( W_i \) and \( \beta_i \) (\( 1 \leq i \leq L \)) such that

\[
\hat{F}(I_i) = O_i, \quad \text{for all } i = 1, 2, ..., k. \tag{2.3}
\]

Since the sigmoid \( S \) is injective and \( \beta_0 \) is assumed to be fixed, the \( IP1 \) can be reduced to an apparently simpler problem that circumvents having to treat the three possible choices for \( T_0 \) separately. This problem is defined next.

Let \( F: \mathbb{R}^{m_0} \rightarrow \mathbb{R}^{m_L} \) denote the map defined by Equation 2.1 with \( T_0 \) equal to the identity map and \( T_L \) replaced by \((\hat{\beta}_L \circ W_L): \mathbb{R}^{m_L} \rightarrow \mathbb{R}^{m_L}\).

**The Interpolation Problem (IP).** Given a set of points of interpolation,

\[
\Omega = \{(x_i, y_i) \in \mathbb{R}^{m_0} \times \mathbb{R}^{m_L}: 1 \leq i \leq k \text{ and } x_i \neq x_j \text{ for } i \neq j\}, \tag{2.4}
\]
determine \( W_i \) and \( \beta_i \) (\( 1 \leq i \leq L \)) such that

\[
F(x_i) = y_i, \quad \text{for all } i = 1, 2, ..., k. \tag{2.5}
\]

The following proposition shows to what extent the two interpolation problems are equivalent.

**Proposition 2.1.** Let the integer \( k > 0 \) be fixed.

(a) If \( IP \) has a solution for every set of interpolation points \( \Omega \) as in Equation 2.4, then \( IP1 \) has a solution for every set of interpolation points \( \hat{\Omega} \) as in Equation 2.2.

(b) If \( IP1 \) has a solution for every set \( \hat{\Omega} \) as in Equation 2.2, then \( IP \) has a solution for every set of interpolation points \( \Omega \), where \( \Omega \) is as in Equation 2.4 if \( T_0 = \text{identity map} \) and \( \Omega = \{(x_i, y_i) \in (-1, 1)^{m_0} \times \mathbb{R}^{m_L}: 1 \leq i \leq k \text{ and } x_i \neq x_j \text{ for } i \neq j\} \), if \( T_0 = S_{m_0} \) or \( T_0 = S_{m_0} \circ \hat{T}_0 \).
Proof. The result follows from the definition of $T_0^{-1}$ and the fact that 
\[ \hat{F}(I) = O \iff F(x) = y \text{ with } x = T_0(I) \text{ and } y = S_m^{-1}(O). \]

It follows from part (a) of the proposition that it suffices to solve the IP. The proposition also shows how to handle the variations in the use of the first layer. Consequently, for the remainder of the paper it will be assumed, without loss of generality, that the FLNN has transfer function $F : \mathbb{R}^{m_0} \to \mathbb{R}^{m_L}$ and the FLNN will be referred to as "the net."

Remark 2.2. One reason for requiring $S$ to be injective is for Proposition 2.1 to hold. This increases the generality of the results that follow by allowing several possible forms for $T_0$ (see Remark 2.1) at the expense of restricting the class of permissible activation functions. Another alternative is to assume that the transfer function of the FLNN is given by $F$ (which is the case for the remainder of the paper) and remove the restriction on $S$ of being injective. If this is the case, then the class of permissible activation functions may include sigmoids that saturate, in which case it will be assumed that the range of $S$ is the closed interval $[-1, 1]$. To avoid confusion, we will indicate when a result or definition also holds for noninjective activation functions. Thus, $S$ is assumed to be injective throughout, unless otherwise stated.

In order to facilitate the statements of some results, we make the following:

Definition 2.1.

(a) We shall say that $F$ interpolates through $\Omega$ if $\Omega$ is as in Equation 2.4 and Equation 2.5 holds.
(b) If IP has a solution, we shall say that $\Omega$ is realizable by the net.
(c) The largest integer $k$ with the property that every set of points $\Omega$ as in Equation 2.4 is realizable by a net with architecture $\tilde{m}$ will be called the interpolation capacity of the net and will be denoted by $\text{IC}(\tilde{m})$.

* The symbol $\square$ indicates the end of a proof.
A related problem is that of characterizing $IC(\tilde{m})$ for all $\tilde{m} \in N^{L+1}$. (Here, $N^{L+1}$ denotes the set of $L + 1$ tuples of positive integers.)

By a simple dimensionality argument, it was shown (Reference 2) that for any continuous activation function $S: R \to [-1, 1]$ and for all $\tilde{m} \in N^{L+1}$ the interpolation capacity $IC(\tilde{m})$ is bounded above by $C(\tilde{m})$, where

$$C(\tilde{m}) = \frac{1}{m_L} \sum_{i=1}^{L} m_i(m_{i-1} + 1) \quad (\tilde{m} \in N^{L+1}). \quad (2.6)$$

Note that the augmented matrices $[W_i; \beta_i]$ are members of $R^{m_i \times (m_{i-1} + 1)}$, $i = 1, 2, \ldots, L$. Thus, $C(\tilde{m})$ is simply the number of degrees of freedom (that is, the number of parameters that need to be specified in order to define $F$ uniquely) divided by the number of outputs.

As a corollary to the next proposition, one can obtain a sharper upper bound for $IC(\tilde{m})$ for all $\tilde{m} \in N^{L+1}$ with $m_L > m_{L-1} + 1$.

**Proposition 2.2.** If $\Omega$ is realizable, then at most $m_{L-1} + 1$ of the vectors $y_i \in R^{m_L}$ $(1 \leq i \leq k)$ can be linearly independent.

We need the following notation in the proof of this proposition: let $F_i = T_i$ and $F_n = T_n \circ F_{n-1}$ for $2 \leq n < L$.

**Proof.** Since $\Omega$ is realizable, there exist $W_i$ and $\beta_i$ $(1 \leq i \leq L)$ such that

$$y_i = F(x_i) = (\hat{W}_L \circ W_L \circ F_{L-1})(x_i) = W_L \circ F_{L-1}(x_i) + \beta_L = [W_L; \beta_L] \left[ \begin{array}{c} \hat{F}_{L-1}(x_i) \\ 1 \end{array} \right] \quad (1 \leq i \leq k). \quad (2.7)$$

Since the set $\left\{ \left[ \begin{array}{c} \hat{F}_{L-1}(x_i) \\ 1 \end{array} \right] \in R^{m_{L-1}+1} : 1 \leq i \leq k \right\}$ contains at most $m_{L-1} + 1$ linearly independent vectors, Equation 2.7 implies that at most $m_{L-1} + 1$ of the vectors $y_i$ $(1 \leq i \leq k)$ can be linearly independent.
Corollary 2.1. If \( m_L > m_{L-1} + 1 \), \( k > m_{L-1} + 1 \), and the set \( (y_i \in \mathbb{R}^{m_L}: 1 \leq i \leq k) \) contains more than \( m_{L-1} + 1 \) linearly independent vectors, then the set \( \{(x_i, y_i) \in \mathbb{R}^{m_0 \times m_L}: 1 \leq i \leq k\} \) is not realizable by a net with architecture \( \hat{m} = (m_0, m_1, ..., m_L) \) for any choice of vector \( x_i \in \mathbb{R}^{m_0} (1 \leq i \leq k) \).

Corollary 2.2. If \( \hat{m} = (m_0, m_1, ..., m_L) \) with \( m_L > m_{L-1} + 1 \), then \( IC(\hat{m}) \leq m_{L-1} + 1 \).

Remark 2.3. Proposition 2.2 and Corollaries 2.1 and 2.2 hold for any activation function \( S \) if the net has transfer function \( F \). Corollaries 2.1 and 2.2 also holds for a net with transfer function \( \hat{F} \) if \( S \) is invertible.

The following example shows that there are architectures \( \hat{m} \) for which \( IC(\hat{m}) < C(\hat{m}) \).

Example 2.1. If \( L = 2, m_0 = m_1 + 1, m_2 = m_1 + 2, \) and \( m_1 \geq 1 \), then \( C(\hat{m}) = 2m_1 + 1 \). Since \( m_2 > m_1 + 1 \), by Corollary 2.2 we have \( IC(\hat{m}) \leq m_1 + 1 \), which is strictly less than \( C(\hat{m}) \).

3. THE LAST LAYER OF WEIGHTS: A LOWER BOUND FOR \( IC(\hat{m}) \)

In this section, we present a characterization of the last layer of weights \( [W_L: \beta_L] \) of a net whose transfer function \( F \) interpolates through a set of points \( \Omega = \{(x_i, y_i): 1 \leq i \leq k\} \). This characterization involves the inverse of the matrix \( X_\alpha(p') \) defined below. Conditions under which \( X_\alpha(p') \) is invertible are explored. We find that \( X_\alpha(p') \) is invertible under very mild conditions, in which case one obtains the following lower bound for \( IC(\hat{m}) \):

\[
IC(\hat{m}) \geq m_{L-1} + 1, \quad \text{for } \hat{m} = (m_0, m_1, ..., m_L). \tag{3.1}
\]

Let \( \tilde{m} = (m_0, m_1, ..., m_L) \in N^{L+1}, P(\tilde{m}) = \prod_{i=1}^{L} \mathbb{R}^{m_i \times (m_{i-1}+1)} \), and \( P'(\tilde{m}) = \prod_{i=1}^{L} \mathbb{R}^{m_i \times (m_{i-1}+1)} \). If a net has architecture \( \tilde{m} \), then the collection of all transfer functions associated with the net for every possible set of weights is clearly parametrized by \( P(\tilde{m}) \). For each point \( p = ([W_1: \beta_1], [W_2: \beta_2], ..., [W_L: \beta_L]) \in P(\tilde{m}) \),
let \( F_p \) denote the transfer function of the net with weights \( p \). Similarly, for each \( p' = ([W_1:β_1], [W_2:β_2], ..., [W_{L-1}:β_{L-1}]) \in P'(m) \), let

\[
F_{p'} = (S_{m_{L-1}} β_{L-1} W_{L-1}) \circ \cdots \circ (S_{m_1} β_1 W_1).
\]

If \( Π: P(m) \rightarrow P'(m) \) denotes the map \( ([W_1:β_1], ..., [W_{L}:β_{L}]) \rightarrow ([W_1:β_1], ..., [W_{L-1}:β_{L-1}]) \), then

\[
F_p = (β_L \circ W_L \circ Π(p)), \text{ for every } p = ([W_1:β_1], ..., [W_{L}:β_{L}]) \in P(m). \quad (3.2)
\]

Let \( Ω = \{(x_i, y_i) : 1 \leq i \leq k\} \) be a set of interpolation points, \( Ω_α = \{(x_{α_1}, y_{α_1}), ..., y_{α_η}) \) where \( α = (α_1, α_2, ..., α_η) \) is a multi-index with \( α_i \in \{1, 2, ..., k\} \) for \( 1 \leq i \leq η \). For each \( p' \in P'(m) \) and multi-index \( α \), define

\[
X_α(p') = \left[ \begin{array}{c}
F_{p'}(x_{α_1}) & F_{p'}(x_{α_2}) & \cdots & F_{p'}(x_{α_η}) \\
1 & 1 & \cdots & 1
\end{array} \right] \in \mathbb{R}^{(m_{L-1} + 1) \times η} \quad (3.3)
\]

and

\[
Y_α = \{y_{α_1}, y_{α_2}, ..., y_{α_η}\} \in \mathbb{R}^{m_0 \times η}.
\]

**Proposition 3.1.** Let \( η = m_{L-1} + 1, k ≥ η \), and \( α = (α_1, α_2, ..., α_η) \) with \( α_i \in \{1, 2, ..., k\} \) for \( 1 \leq i \leq η \).

(a) If \( p \in P(m) \), \( X_α(Π(p)) \) is invertible, and \( F_p \) interpolates through \( Ω_α \), then \( p = (Π(p), [W_{L}:β_{L}]) \) with

\[
[W_{L}:β_{L}] = Y_α \left(X_α(Π(p))^{-1}\right). \quad (3.4)
\]

(b) Conversely, if \( p' \in P'(m) \), \( X_α(p') \) is invertible, Equation 3.4 holds, and \( p = (p', [W_{L}:β_{L}]) \), then \( F_p \) interpolates through \( Ω_α \).

**Proof.** It follows from Equation 2.7 that \( F_p \) interpolates through \( Ω_α \) if, and only if, \( p = (Π(p), [W_{L}:β_{L}]) \) and

\[
Y_α = [W_{L}:β_{L}] X_α(Π(p)). \quad (3.5)
\]
Thus, if $X_\alpha(\Pi(p))$ is invertible, then Equation 3.5 implies Equation 3.4. This proves (a). Conversely, if Equation 3.4 holds, then $[W_L; \beta_L]$ satisfies Equation 3.5 which proves (b).

**Remark 3.1.** If $\mu < \eta = m_{L-1} + 1$, $\alpha = (\alpha_1, \alpha_2, ..., \alpha_\mu)$ and the columns of $X_\alpha(p')$ are linearly independent, then there exists a matrix $U \in \mathbb{R}^{\eta \times (\eta - \mu)}$ such that the augmented matrix $[X_\alpha(p') : U] \in \mathbb{R}^{\eta \times \eta}$ is invertible. If the inverse of $[X_\alpha(p') : U]$ is partitioned as

$$
\begin{bmatrix}
V_1 \\
V_2
\end{bmatrix}
$$

with $V_1 \in \mathbb{R}^{\mu \times \eta}$ and $V_2 \in \mathbb{R}^{(\eta - \mu) \times \eta}$, then $V_1 X_\alpha(p') = I_\mu$, where $I_\mu$ is the $\mu \times \mu$ identity matrix. Therefore, if $[W_L; \beta_L] = Y_\alpha V_1$, then $[W_L; \beta_L]$ satisfies Equation 3.5 so that with $p = (p', Y_\alpha V_1)$, $F_p$ interpolates through $\Omega_\alpha$.

For each multi-index $\alpha = (\alpha_1, \alpha_2, ..., \alpha_\eta)$, with $\eta = m_{L-1} + 1$, let $E_\alpha$ denote the set of all points $p'$ in $P'(\vec{m})$ such that the matrix $X_\alpha(p')$ is invertible. We may define a map $\Gamma_\alpha: E_\alpha \rightarrow P(m)$ by $\Gamma_\alpha(p') = (p', Y_\alpha[X_\alpha(p')]^{-1})$. Part (b) of Proposition 3.1 says that all the transfer functions $F_{\Gamma_\alpha(p')}(p' \in E_\alpha)$ interpolate through $\Omega_\alpha$; that is, they satisfy

$$
F_{\Gamma_\alpha(p')}(x) = y \quad \text{for all } (x, y) \in \Omega_\alpha. \quad (3.6)
$$

Consequently, if $E_\alpha$ is not empty, then there exists $p \in P(\vec{m})$ such that $F_p$ interpolates through $\Omega_\alpha$.

For a given multi-index $\alpha$, whether or not the set $E_\alpha$ is empty is difficult to answer in general, since it depends on the type of activation function in the net. The next proposition shows that under certain conditions the set $E_\alpha$ is "large" (see Remark 3.2). In order to state the proposition, we must introduce the map $\Lambda_\alpha: P'(\vec{m}) \rightarrow \mathbb{R}$ defined by $\Lambda_\alpha(p) = \det X_\alpha(p)$, where $X_\alpha(p)$ is given by Equation 3.3, $\eta = m_{L-1} + 1$ and $\det X_\alpha(p)$ denotes the determinant of $X_\alpha(p)$. Let

$$
\nabla \Lambda_\alpha: P'(\vec{m}) \rightarrow \mathbb{R}^\xi
$$

denote the gradient of $\Lambda_\alpha$, where $\xi = \sum_{i=1}^{L-1} m_i (m_{i-1} + 1)$. A point $p$
in $P'(\mathbb{m})$ has $\xi$ coordinates and will be denoted by $p = (p_1, p_2, \ldots, p_\xi)$; that is, we will identify $P'(\mathbb{m})$ with $\mathbb{R}^\xi$.

If $E_\alpha^c$ denotes the complement of $E_\alpha$ in $P'(\mathbb{m})$, then $E_\alpha^c = \Lambda_\alpha^{-1}(0)$. Hence, $E_\alpha^c$ is closed. Let $E'_\alpha$ be the set of points $p$ such that $\nabla \Lambda_\alpha(p) = 0$.

**Proposition 3.3.** If the activation function $S: \mathbb{R} \to (-1, 1)$ is continuously differentiable, then for each multi-index $\alpha$ with $\eta$ components the set $E_\alpha^c - E'_\alpha$ has Lebesgue measure zero.

The proof of this proposition is based on the following.

**Lemma 3.1.** If $S$ is continuously differentiable, then for each $p \in E_\alpha^c - E'_\alpha$ there exists an open set $O$ in $P'(\mathbb{m})$ such that $p \in O$ and $m(E_\alpha^c \cap O) = 0$, where $m$ denotes the Lebesgue measure on $P'(\mathbb{m})$.

**Proof:** If $p^0 = (p_1^0, p_2^0, \ldots, p_\xi^0) \in E_\alpha^c$ and $\nabla \Lambda_\alpha(p^0) \neq 0$, then $\frac{\partial \Lambda_\alpha}{\partial p_j}(p^0) \neq 0$ for some $j \in \{1, 2, \ldots, \xi\}$. To simplify the notation, we may assume that $j = 1$. Write $p^0 = (p_1^0, q^0)$ with $q^0 = (p_2^0, p_3^0, \ldots, p_\xi^0)$. Since the activation function $S$ is continuously differentiable, so is $\Lambda_\alpha$; therefore, by the Implicit Function Theorem (Reference 3), there exist open sets $V \subset \mathbb{R}^1$ and $U \subset \mathbb{R}^{\xi-1}$ with $(p_1^0, q^0) \in V \times U$ and a unique map $\varphi: U \to V$ such that $\varphi(q^0) = p_1^0$, $\Lambda_\alpha(\varphi(q), q) = 0$ for all $q \in U$, and $\Lambda_\alpha(p_1, q) \neq 0$ if $(p_1, q) \in V \times U$ and $p_1 \neq \varphi(q)$. Thus,

$$E_\alpha^c \cap [V \times U] = \{ (\varphi(q), q) : q \in U \}. \tag{3.7}$$

Next we will show that $m(E_\alpha^c \cap [V \times U]) = 0$. 


Let $m_i$ denote the Lebesgue measure on $\mathbb{R}^l$ ($l = 1, 2, ...$). Since $E^c_\alpha$ is closed, $Q = E^c_\alpha \cap (V \times U)$ is a Borel set in $\mathbb{R}^5$. Therefore, by the definition of the product measure $m_1 \times m_{\xi-1}$ (Reference 4) and Theorem 7.11 of Reference 4, we have

$$m_\xi(Q) = (m_1 \times m_{\xi-1})(Q) = \int_U m_1(\{p_1 : (p_1, q) \in Q\}) \ dm_{\xi-1}(q).$$

This together with Equation 3.7 gives

$$m_\xi(Q) = \int_U m_1(\{q\}) \ dm_{\xi-1}(q) = \int_U 0 \ dm_{\xi-1}(q) = 0.$$

By letting $O = V \times U$, the proof is finished. \\

Proof of Proposition 3.3. By Lemma 3.1, for every $p \in E^c_\alpha - E^*_{\alpha'}$, there exists an open set $O_p$ such that $p \in O_p$ and $m(E^c_\alpha \cap O_p) = 0$. Since every Euclidean space is second countable (Reference 5), $E^c_\alpha - E^*_{\alpha'}$ can be covered by a countable collection of the sets $O_p$. Finally, since Lebesgue measure is countably additive, we conclude that $m(E^c_\alpha - E^*_{\alpha'}) = 0$. \\

Remark 3.2. If also $m(E^*_{\alpha'} \cap E^c_\alpha) = 0$, then it follows from Proposition 3.3 that all the matrices $X_\alpha(p)$ ($p \in P'(\tilde{m})$) are invertible except for those $p$ in the set $E^c_\alpha$ of measure zero.

As a corollary to Lemma 3.1, we can obtain a weaker condition for the existence of an invertible $X_\alpha(p)$.

Corollary 3.1. If $\nabla \Lambda_\alpha$ is not identically zero, then there exists $p \in P'(\tilde{m})$ such that $X_\alpha(p)$ is invertible; that is $E_\alpha \neq \emptyset$. 

13
Proof. If $\nabla \Lambda_\alpha$ is not identically zero, there is a point $p \in P'(\tilde{m})$ such that $\nabla \Lambda_\alpha(p) \neq 0$. Either $\Lambda_\alpha(p) = 0$ or not. If $\Lambda_\alpha(p) \neq 0$, then $X_\alpha(p)$ is invertible. If $\Lambda_\alpha(p) = 0$, then $p \in E_\alpha^c - E_\alpha$ and by Lemma 3.1 there is an open set $O$ in $P'(\tilde{m})$ such that $p \in O$ and $m(E_\alpha^c \cap O) = 0$. The set $O$ must have positive measure, since it is open and nonempty. Consequently, $O \cap E_\alpha$ is not empty (otherwise, $O \subseteq S_\alpha^c$—a contradiction) and any of its members satisfy the conclusion of the corollary.

Corollary 3.2. If $\nabla \Lambda_\alpha$ is not identically zero, then there exists $p \in P'(\tilde{m})$ such that $F_{\Gamma_\alpha(p)}$ interpolates through $\Omega_\alpha$.

Proof. This follows from Corollary 3.1 and Equation 3.6.

4. NETS WITH TWO LAYERS OF WEIGHTS

In this section, we will show how to define the weights of a net with two layers of weights so that the transfer function of the net interpolates through a realizable set $\Omega = \{(x_i, y_i) : 1 \leq i \leq k\}$ when the matrix $X = \begin{bmatrix} x_1; x_2; \ldots; x_k \\ 1; 1; \ldots; 1 \end{bmatrix}$ has rank $k$, where $k \leq m_0 + 1$.

Assume that $L = 2$, so that $\tilde{m} = (m_0, m_1, m_2)$ and assume $m_2 \leq m_1$.

Choose a maximal subset of $\{y_i : 1 \leq i \leq k\}$ consisting of linearly independent vectors and let $\eta$ be its cardinality; $\eta \leq m_2$. Without loss of generality, we may assume that $y_1, y_2, \ldots, y_\eta$ are linearly independent (the vectors $y_i$ may be relabeled if necessary). There exist constants $a_{ij} \in \mathbb{R}$ such that

$$y_i = \sum_{j=1}^{\eta} a_{ij}y_j, \quad \eta + 1 \leq i \leq k.$$  \hspace{1cm} (4.1)

Set $\hat{\alpha} = \max\{|a_{ij}| : 1 \leq j \leq \eta, \eta + 1 \leq i \leq k\}$.  

14
If the matrix $X$ defined above has rank $k$, then there exists a matrix $V \in \mathbb{R}^{k \times k_0}$ such that $ VX = I_k $, where $I_k$ is the $k \times k$ identity matrix (see the argument in Remark 3.1). If $k = m_0 + 1$, then $V = X^{-1}$, the inverse of $X$. Let $e_i^{m_1}$ denote the $i^{th}$ column of the $m_1 \times m_1$ identity matrix ($1 \leq i \leq m_1$). Fix a number $\varepsilon \in (0, 1)$ such that $\bar{e} < 1$ and consider the following vectors $z_i \in \mathbb{R}^{m_1}, 1 \leq i \leq k$.

$$ z_i = S^{-1}_{m_1}(\varepsilon e_i^{m_1}) = S^{-1}(\varepsilon) e_i^{m_1}, \quad \text{for } 1 \leq i \leq \eta \quad (4.2a) $$

$$ z_i = S^{-1}_{m_1} \left( \sum_{j=1}^{\eta} e_{aij} e_j^{m_1} \right) = \sum_{j=1}^{\eta} S^{-1}(e_{aij}) e_j^{m_1}, \quad \text{for } \eta + 1 \leq i \leq k. \quad (4.2b) $$

Define an $m_1 \times k$ matrix $Z$ by $Z = [z_1 : z_2 : \ldots : z_k]$, and set

$$ [W_1 : \beta_1] = ZV, \quad (4.3) $$

$$ W_2 = \frac{1}{\varepsilon} [y_1 : y_2 : \ldots : y_\eta : 0 : \ldots : 0] \in \mathbb{R}^{m_2 \times m_1}, \quad \text{and } \beta_2 = 0. \quad (4.4) $$

**Proposition 4.1.** Assume $X$ has rank $k$. Using the notation of Section 3 with $L = 2$, if $p = ([W_1 : \beta_1], [W_2 : \beta_2])$ and Equations 4.3 and 4.4 hold, then $F_p$ interpolates through $\Omega$.

**Proof:** Since $ VX = I_k $, it follows that $ V \begin{bmatrix} x_i \\ 1 \end{bmatrix} = e_i^k $, where $e_i^k$ is the $i^{th}$ column of the $k \times k$ identity matrix, $1 \leq i \leq k$. Thus, by Equation 4.3,

$$ [W_1 : \beta_1] \begin{bmatrix} x_i \\ 1 \end{bmatrix} = ZV \begin{bmatrix} x_i \\ 1 \end{bmatrix} = Z e_i^k = z_i \quad \text{for } 1 \leq i \leq k. $$

Consequently, by Equations 4.2a and 4.4, for $1 \leq i \leq \eta$, we have

$$ F_p(x_i) = [W_2 : \beta_2] \begin{bmatrix} S_{m_1}(z_i) \\ 1 \end{bmatrix} = W_2(\varepsilon e_i^{m_1}) = y_i. $$

Now, by Equations 4.1, 4.2b, and 4.4, for $\eta + 1 \leq i \leq k$, we have
\[ F_p(x_i) = \begin{bmatrix} W_2 \beta_2 \end{bmatrix} = W_2 \left( \sum_{j=1}^{n} \varepsilon a_{ij} e_j^m \right) = \sum_{j=1}^{n} a_{ij} y_j = y_i. \]

Hence, \( F_p \) interpolates through \( \Omega \).

5. COMPARISONS AND APPLICATIONS

Two different techniques for determining the weights of a FLNN so that its transfer function interpolates through a given set of points \( \Omega \) have been presented. If the cardinality of \( \Omega \) is \( k \), then both techniques require a certain matrix to have rank \( k \) in order to interpolate through all of \( \Omega \).

The technique presented in Section 3 (Technique 1) requires the columns of \( X_{k}(p') \) to be linearly independent \((k = (1, 2, \ldots, k))\); see Remark 3.1. Since the columns of \( X_{k}(p') \) are \((m_{L-1} + 1)\)-dimensional, \( k \) can be at the most \( m_{L-1} + 1 \). Thus, using Technique 1, the net can interpolate through an arbitrary set \( \Omega \) as in Equation 2.4 with at most \( m_{L-1} + 1 \) points, provided \( \nabla A_k \) is not identically zero (see Corollary 3.1).

The technique presented in Section 4 (Technique 2) was developed for nets with \( \hat{m} = (m_0, m_1, m_2) \) and \( m_2 \leq m_1 \). It requires the columns of \( X \) to be linearly independent. Since the columns of \( X \) are \((m_0 + 1)\)-dimensional, \( k \) can be at the most \( m_0 + 1 \). Thus, using Technique 2, the net can interpolate through a set \( \Omega \) as in Equation 2.4 with at most \( m_0 + 1 \) points, provided the matrix \( X \) has full rank.

Aside from the difference in the number of points through which the net can interpolate using the two techniques, there is another difference. Technique 2 will specify all the weights in the net, while Technique 1 only specifies the last layer of weights in terms of the first \( L - 1 \) layers of weights. When the activation function \( S \) is such that the gradient of \( A_k \) vanishes only on a set of measure zero, then as Remark 3.2 suggests, one may choose the first
L - 1 layers of weights at random and, loosely speaking, with probability 1 a nonsingular matrix $X_k(p')$ is obtained when $k = m_{L-1} + 1$. More precisely, $X_k(p')$ will be nonsingular for all $p'$ except for some points $p'$ in a set of Lebesgue measure zero. It can be shown that any activation function $S$, which becomes an entire function (Reference 6) when extended to the complex plane, has the property that $\nabla A_k$ is zero only on a set of measure zero unless it vanishes identically, in which case $A_k$ itself is identically zero. An example of such an $S$ is $S(t) = 2f(t) - 1 \ (t \in \mathbb{R})$, where

$$f(t) = \frac{1}{\sqrt{n}} \int_{-\infty}^{\infty} e^{-x^2} \, dx, \quad (t \in \mathbb{R}).$$

In a similar vein, when $k = m_0 + 1$, the matrix $X$ will be nonsingular with probability 1. That is, $X$ is of the form $\begin{bmatrix} \ldots & A & \ldots \\ 1 & 1 & \ldots & 1 \end{bmatrix}$ with $A \in \mathbb{R}^{m_0 \times k}$, and the set of matrices $A$ in $\mathbb{R}^{m_0 \times k}$ for which $X$ is singular has Lebesgue measure zero.

For nets with only one hidden layer ($L = 2$) and $m_2 \leq m_1$, one has a choice of interpolation techniques. If the goal is to interpolate through as many points as possible, clearly one uses the technique with the largest permissible $k$. However, usually one can select $m_1$; thus, by choosing $m_1$ large enough, one can interpolate through any number of points using Technique 1. On the other hand, in some applications the input layer may be large enough already; in which case, Technique 2 may be adequate with a more conservative value for $m_1$.

Perhaps these techniques for determining the weight will prove to be most useful in the initialization of weights. Some of the most popular learning algorithms in use today (e.g., back propagation (Reference 1)) are based on iterative steepest descent minimization procedures, where, at each step, the approximate solution is corrected in the direction of steepest descent in order to reduce the error. To begin an initial set of weights is required, which is usually chosen at random. The speed of convergence depends very heavily on the quality of the initial weights; that is, on how close the initial weights are to
the correct set of weights. To improve the quality of the initial weights, one could select \( m_0 + 1 \) representative input-output pairs (prototypes) such that the matrix \( X \) is invertible and calculate an initial set of weights using Technique 2. Alternatively, one could select \( m_{L-1} + 1 \) representative input-output pairs, choose the first \( L - 1 \) layer of weights at random, and calculate the last set of weights using Technique 1.

6. JACOBIAN MATRIX OF THE TRANSFER FUNCTION

Since a small input change to the net produces an output change whose magnitude is approximately bounded by the product of the operator norm of the Jacobian matrix (total derivative) of the transfer function and the magnitude of the input change, we suggest that the sensitivity of the transfer function \( F \) to noisy input patterns at the interpolation points can be measured by the norm of the Jacobian matrix of \( F \) at the interpolation points. Thus, in this section, we derive an expression for the Jacobian matrix of \( F \) and compute an upper bound for its norm.

Let \( W^l_{ij} (1 \leq i \leq m_l, 1 \leq j \leq m_{l-1}) \) denote the entries of the weight matrix \( W^l (1 \leq l \leq L) \), and let \( \beta^l_j (1 \leq j \leq m_l) \) denote the components of the vector \( \beta^l (1 \leq l \leq L) \).

For each \( z \in \mathbb{R}^{m_{l-1}} \) with components \( z_i (1 \leq i \leq m_{l-1}) \), let \( D(z) \) be an \( m_l \times m_l \) diagonal matrix defined by
where \( S' \) denotes the derivative of \( S \). The expression above defines a map
\[
D_l : \mathbb{R}^{m \times 1} \rightarrow \mathbb{R}^{m \times m} \quad (1 \leq l \leq L).
\]
Recall that \( F = (\Phi L_0 W_L )^o T_{L-1} \circ \cdots \circ T_1 \) and \( F_n = T_n \circ F_{n-1} \)
for \( 1 \leq n < L \) with \( F_0 \) the identity map on \( \mathbb{R}^{m \times 0} \). Let \( T_l(z) \) denote the Jacobian
matrix of \( T_l \) at \( z \in \mathbb{R}^{m \times 1} \) and \([T_l(z)]_{ij}\) its \( ij \)-th entry \((1 \leq i \leq m, 1 \leq j \leq m_{l-1}, 1 \leq l < L)\).
Use a similar notation for the Jacobian matrices of \( F \) and \( F_n \) \((1 \leq n < L)\).

**Proposition 6.1.** Assume \( L \geq 2 \).

\[
T_l(z) = D_l(z) W_l \quad (z \in \mathbb{R}^{m \times 1}, 1 \leq l < L). \tag{6.1}
\]

\[
F'(x) = W_L \prod_{l=1}^{L-1} [D_l(F_{l-1}(x)) W_l] \quad (x \in \mathbb{R}^{m \times 0}). \tag{6.2}
\]

**Proof:** If the components of \( z \in \mathbb{R}^{m \times 1} \) are \( z_i \) \((1 \leq i \leq m_{l-1})\), then

\[
T_l(z) = \begin{bmatrix}
S' \left( \sum_{j=1}^{m_{l-1}} W_{ij} z_j + \beta'_1 \right),
S' \left( \sum_{j=1}^{m_{l-1}} W_{2j} z_j + \beta'_2 \right),
\ldots,
S' \left( \sum_{j=1}^{m_{l-1}} W_{m_{l-1}j} z_j + \beta'_{m_{l-1}} \right)
\end{bmatrix}^T.
\]

Consequently,
\[
[T_l(z)]_{ij} = \frac{\partial}{\partial z_j} S' \left( \sum_{n=1}^{m_{l-1}} W_{1n} z_n + \beta'_i \right) = S' \left( \sum_{n=1}^{m_{l-1}} W_{1n} z_n + \beta'_i \right) W_{ij}. \quad (1 \leq j \leq m_{l-1}, 1 \leq i \leq m_l, 1 \leq l < L).
\]

This gives Equation 6.1.
Applying the chain rule to the composition $T_n \circ F_{n-1}$ and using Equation 6.1, one obtains for $n = 1, 2, ..., L - 1$

$$F'_n(x) = T'_n(F_{n-1}(x))F'_{n-1}(x) = D_n(F_{n-1}(x))W_n F'_{n-1}(x) \quad (x \in \mathbb{R}^{m_0}).$$

Hence, by induction we have

$$F'_{L-1}(x) = \prod_{i=L-1}^{1} [D_i(F_{i-1}(x))W_i] \quad (x \in \mathbb{R}^{m_0}). \quad (6.3)$$

Finally, since $F(x) = W_L F_{L-1}(x) + \beta_L$, $F'(x) = W_L F'_{L-1}(x)$, thus Equation 6.3 implies Equation 6.2.

It should be noted that the product indicated in Equations 6.2 and 6.3 is a product of matrices that may not commute; thus, it is important to understand the correct order of multiplication; namely,

$$[D_{L-1}(F_{L-2}(x))W_{L-1}] \cdot [D_{L-2}(F_{L-3}(x))W_{L-2}] \cdots [D_1(x)W_1].$$

The sensitivity of the transfer function $F$ at a point $x$ will be measured here by the induced $(p, q)$-norm of the linear transformation $F'(x)$ for particular values of $p$ and $q$. The induced $(p, q)$-norms are defined below. Other operator norms could be used; however, the induced $(p, q)$-norms lead to upper bounds for the norm of $F'(x)$ that are computable for certain values of $p$ and $q$ and can be interpreted qualitatively.

The induced $(p, q)$ matrix norms are defined next. A list of properties of $(p, q)$-norms that will be needed is also included. To the author's knowledge, some of the properties of $(p, q)$-norms needed here are not available in the open literature. For completeness, this material is developed in Proposition 6.2.
For $1 \leq p < \infty$, the $p$-norm of a vector $x \in \mathbb{R}^n$ is defined by

$$
\|x\|_p = \left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p},
$$

where $x_i$ ($1 \leq i \leq n$) are the components of $x$ with respect to the standard basis on $\mathbb{R}^n$ consisting of the columns of the $n \times n$ identity matrix $I_n$. If $p = \infty$, then

$$
\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|.
$$

If $A : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation, then the induced $(p, q)$-norm of $A$ is defined by

$$
\|A\|_{pq} = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_q},
$$

where $1 \leq p \leq \infty$ and $1 \leq q \leq \infty$.

Recall that we are using the same symbol to represent a linear transformation and the matrix associated with it with respect to the standard basis on its domain and range. If $A$ is an $m \times n$ matrix, let $A_{ij}$ denote its $ij$th entry ($1 \leq i \leq m$, $1 \leq j \leq n$).

**Proposition 6.2.** Let $A : \mathbb{R}^n \to \mathbb{R}^m$, $B : \mathbb{R}^m \to \mathbb{R}^l$, and $C : \mathbb{R}^l \to \mathbb{R}^m$.

$$
\|BA\|_{pq} \leq \|B\|_p \|A\|_{rq} \quad 1 \leq p, q, r \leq \infty.
$$

(6.4a)

$$
\|A\|_{pq} \leq \|[A;C]\|_{pq} \quad 1 \leq p, q \leq \infty.
$$

(6.4b)

$$
\|A\|_{\infty,1} = \max_{i,j} |A_{ij}|.
$$

(6.5)

If $A$ is a square diagonal matrix, $A = \text{diag}[A_{ii}]$, then

$$
\|A\|_{pq} = \left( \sum_{i=1}^{n} |A_{ii}|^p \right)^{1/p},
$$

$1 \leq p < \infty,$

(6.6)

$$
\|A\|_{p1} = \max_{1 \leq j \leq n} \left( \sum_{i=1}^{m} |A_{ij}|^p \right)^{1/p},
$$

$1 \leq p < \infty,$

(6.7)

$$
\|A\|_{1\infty} = \max_{1 \leq i \leq n} \left( \sum_{j=1}^{m} |A_{ij}| \right),
$$

(6.8)
\[ \|\mathbf{A}\|_p = \max_{1 \leq i \leq m} \left( \sum_{j=1}^{n} |A_{ij}|^{p^*} \right)^{1/p^*} \quad 1 < p < \infty, \tag{6.9} \]

where \( p^* \) is the conjugate exponent to \( p \); that is, \( p^* \) satisfies \( \frac{1}{p} + \frac{1}{p^*} = 1 \).

**Proof:** The definition of \( \|B\|_{pr} \) implies \( \|By\|_p \leq \|B\|_{pr} \|y\|_r \) for all \( y \in \mathbb{R}^m \).

Therefore, \( \|BA\|_{pq} = \sup_{x \neq 0} \frac{|BAx|_p}{|x|_q} \leq \sup_{x \neq 0} \frac{|B|_{pr} |Ax|_r}{|x|_q} = \|B\|_{pr} \|A\|_{rq} \). This gives Inequality 6.4a.

There exists \( x \in \mathbb{R}^n \) such that \( \|x\|_q = 1 \) and \( \|A\|_{pq} = \|Ax\|_p \). Consider the vector \( y = [x_0] \in \mathbb{R}^{n+1} \). Clearly, \( \|y\|_q = \|x\|_q = 1 \) and \( \|[A:C]y\|_p = \|Ax\|_p = \|A\|_{pq} \). Consequently, Inequality 6.4b holds.

If \( \|x\|_1 = 1 \), then \( \|Ax\|_\infty = \max_{1 \leq i \leq m} \left( \sum_{j=1}^{n} A_{ij} x_j \right) \leq \max_{1 \leq i \leq m} \sum_{j=1}^{n} |A_{ij}| x_j \leq \max_{ij} |A_{ij}| \sum_{j=1}^{n} |x_j| = \max_{ij} |A_{ij}|. \) Hence, the left-hand side (LHS) of Equation 6.5 cannot exceed the right-hand side (RHS). If \( |A_{i0}| = \max_{ij} |A_{ij}| \) and \( x \) is the \( j_0 \)th column of \( I_n \), then

\[ \|x\|_1 = 1 \text{ and } \|Ax\|_\infty = \max_{1 \leq i \leq m} \left( \sum_{j=1}^{n} A_{ij} x_j \right) = \max_{1 \leq i \leq m} |A_{ij}| = |A_{i0}|. \] This establishes Equation 6.5.

If \( A \) is a diagonal (square) matrix and \( \|x\|_\infty = 1 \), then

\[ \|Ax\|_p = \left[ \sum_{i=1}^{n} |A_{ii} x_i|^p \right]^{1/p} = \left[ \sum_{i=1}^{n} |A_{ii}|^p |x_i|^p \right]^{1/p} \leq \left[ \sum_{i=1}^{n} |A_{ii}|^p \right]^{1/p}. \]
Consequently, the LHS of Equation 6.6 cannot exceed the RHS. Moreover, if
\[ x = [1, 1, ..., 1]^T \in \mathbb{R}^n, \] then \( \|x\|_\infty = 1 \) and \( \|Ax\|_p = \left[ \sum_{i=1}^n |A_{ij}|^p \right]^{1/p} \). This establishes Equation 6.6.

Let \( 1 \leq p < \infty, \|x\|_1 = 1 \), and consider the m-vectors \( v_j = x_j a_j \), where \( a_j \) denotes the \( j \)th column of \( A \) (\( 1 \leq j \leq n \)). Since the vector norm \( \|\cdot\|_p \) satisfies the triangle inequality, it follows that \( \left\| \sum_{j=1}^n v_j \right\|_p \leq \sum_{j=1}^n \|v_j\|_p \). Consequently,

\[
\|Ax\|_p = \left\{ \sum_{j=1}^m \left\| \sum_{i=1}^n A_{ij} x_j \right\|_p \right\}^{1/p} = \sum_{i=1}^n \left\| \sum_{j=1}^m A_{ij} x_j \right\|_p^{1/p} \leq \sum_{i=1}^n \left\| \sum_{j=1}^m A_{ij} x_j \right\|_p^{1/p} \leq \sum_{i=1}^n \left[ \max_{1 \leq j \leq n} \sum_{i=1}^m |A_{ij}|^p \right]^{1/p} \leq \max_{1 \leq j \leq n} \left[ \sum_{i=1}^m |A_{ij}|^p \right]^{1/p}.
\]

This shows that the LHS of Equation 6.7 cannot exceed the RHS. Now, if the maximum on the RHS of Equation 6.7 occurs at \( j = j_0 \), let \( x \) be the \( j_0 \)th column of the identity matrix. Then \( \|x\|_1 = 1 \) and \( \|Ax\|_p = \sum_{i=1}^m |A_{ij_0}|^p = \max_{1 \leq j \leq n} \sum_{i=1}^m |A_{ij}|^p \). Hence, Equation 6.7 holds.

Equation 6.8 is proved in Reference 7.

Let \( 1 < p < \infty, \|x\|_p = 1 \) and let \( p^* \) be the conjugate exponent to \( p \). Note that \( p(p^* - 1) = p^* \). Then,

\[
\|Ax\|_\infty = \max_{1 \leq i \leq m} \left\{ \sum_{j=1}^n |A_{ij} x_j| \right\} \leq \max_{1 \leq i \leq m} \left\{ \sum_{j=1}^n |A_{ij}| |x_j| \right\} \leq \|x\|_p \cdot \max_{1 \leq i \leq m} \left[ \sum_{j=1}^n |A_{ij}|^p \right]^{1/p^*}.
\]

23
where the second inequality follows from Hölder's inequality (Reference 4). Thus, the LHS of Equation 6.9 cannot exceed the RHS. Now, if the maximum on the RHS of Equation 6.9 occurs at \(i = i_0\), let \(x_j = \alpha A_{i_0 j} |A_{i_0 j}|^{p^* - 2}\) if \(A_{i_0 j} \neq 0\) and \(x_j = 0\) if \(A_{i_0 j} = 0\) \((1 \leq j \leq n)\), where \(\alpha = \left[\sum_{j=1}^{n} |A_{i_0 j}|^{p^*}\right]^{-1/p}\). Since \(p(p^* - 1) = p^*\), the choice of \(\alpha\) gives \(\|x\|_p = 1\) and \(\|\mathbf{A}\|_{ap} \geq |\sum_{j=1}^{n} A_{i_0 j} x_j| = \alpha \sum_{j=1}^{n} |A_{i_0 j}|^{p^*} = \left[\sum_{j=1}^{n} |A_{i_0 j}|^{p^*}\right]^{1/p^*}$. This establishes Equation 6.9.

Remark 6.1. If \(\sigma_{\text{max}}(A)\) and \(\sigma_{\text{min}}(A)\) denote, respectively, the largest and the smallest singular values of the matrix \(A\), then

\[
\|\mathbf{A}\|_2 = \sigma_{\text{max}}(A) \quad \text{(6.10a)}
\]
\[
\|\mathbf{A}^{-1}\|_2 = 1/\sigma_{\text{min}}(A) \quad \text{if } A \text{ is invertible.} \quad \text{(6.10b)}
\]

See References 7 and 8.

Applying Inequality 6.4a to Equation 6.2, one obtains the following upper bound for \(\|F'(x)\|_{pq}\) \((1 \leq p, q \leq \infty, \ (L > 2))\).

\[
\|F'(x)\|_{pq} \leq \|W_L\|_{pq} \|D_{L-1}(F_{L-2}(x))\|_{q-\infty} \|W_{L-1}\|_{1-q} \prod_{l=2}^{L-2} \|D_l(F_{l-1}(x))\|_{1-q} \|W_l\|_{1-q}. \quad \text{(6.11)}
\]

Note that all of the norms appearing in Inequality 6.11 can be computed using the formulas in Proposition 6.2 whenever \((p, q) = (\infty, q), \ 1 \leq q \leq \infty, \text{ or } (p, q) = (p, 1), \ 1 \leq p \leq \infty\). They also can be computed when \((p, q) = (2, 2)\). In particular, if \(\hat{W}^l = \max_{i,j} |W_{ij}^l| \ (1 \leq l \leq L)\) and \([F_{l-1}(x)]_j\) denotes the \(j^{\text{th}}\) component of \(F_{l-1}(x)\) \((1 \leq j \leq m_{l-1}, \ 1 \leq l \leq L-1)\), then when \((p, q) = (\infty, 1)\), one obtains

\[
\|F'(x)\|_{1-q} \leq \prod_{l=1}^{L} \hat{W}^l \cdot \prod_{l=1}^{L-1} \sum_{j=1}^{m_l} S \left( \sum_{j=1}^{m_{l-1}} W_{ij}^l [F_{l-1}(x)]_j + \beta^l \right). \quad \text{(6.12)}
\]
Inequality 6.12 is interesting, since it shows that, qualitatively speaking, the sensitivity at a point \((x, F(x))\) will be small if the weights are small. In particular, if the derivative of the activation function \(S\) is bounded, say \(S'(t) \leq \eta\) for all \(t \in \mathbb{R}\), then

\[
\|F'(x)\|_{\infty} \leq \eta^{L-1} \prod_{i=1}^{L} \hat{W}_i \cdot \prod_{i=1}^{L} m_i\quad (x \in \mathbb{R}^m).
\]

Moreover, if \(S'(t) \to 0\) as \(t \to \pm \infty\), then Inequality 6.12 suggests that the sensitivity (i.e., the norm of the Jacobian matrix) can be made small at the points of interpolation by choosing \(|\beta_i|\) large (1 \(\leq i \leq m_1\), 1 \(\leq l \leq L - 1\), \(L \geq 2\)). Whether one can \textit{simultaneously} choose the biases \(|\beta_i|\) large, keep the weights small, and interpolate through a set of points \(\Omega\) is a topic for further research.

As an example of the applications of the theory developed in this section, we can investigate the sensitivity of the transfer functions that are obtained using the Interpolation Techniques 1 and 2 that were presented in earlier sections. It will be assumed that \(L = 2\).

When \(L = 2\), Proposition 6.1 gives

\[
F'(x) = W_2 D_1(x) W_1\quad (x \in \mathbb{R}^m)
\]

from which it follows (by Inequality 6.4a),

\[
\|F'(x)\|_{pq} \leq \|W_2\|_{pr} \|D_1(x)\|_{rq} \|W_1\|_{lq}.\quad (6.13)
\]

If \(t = \infty\), we can use Equation 6.6 to compute \(\|D_1(x)\|_{l\infty}\) for \(1 \leq r < \infty\). Note that the elements of the diagonal matrix \(D_1(x)\) coincide with the components of the vector \(S_{m_1}(\{W_1: \beta_1\}^T_{1 \times 1})\), where \(S_{m_1}\) is the map defined by \(S_{m_1}(z) = [S'(z_1), S'(z_2), ..., S'(z_{m_1})]^T\) for every \(z = [z_1, z_2, ..., z_{m_1}]^T \in \mathbb{R}^{m_1}\). Therefore, it follows from Equation 6.6 that the \((r, \infty)\)-norm of the matrix \(D_1(x)\) coincides with the \(r\)-norm of the vector \(S_{m_1}(\{W_1: \beta_1\}^T_{1 \times 1})\).
\begin{equation}
\|D_1(x)\|_w = \|S'_{m_1}(\left[ \begin{array}{c} x \\ 1 \end{array} \right])\|_r \quad (x \in \mathbb{R}^{m_0}). \tag{6.14}
\end{equation}

Now, consider Technique 2 for interpolating through \( \Omega = \{(x_1, y_1) \in \mathbb{R}^{m_0} \times \mathbb{R}^{m_2}; 1 \leq i \leq k\} \), with \( k = m_0 + 1 \) and \( m_2 \leq m_1 \). Assume \( X \) is invertible. Recall that

\[
[W_1; \beta_1] \left[ \begin{array}{c} x_i \\ 1 \end{array} \right] = z_i = \sum_{j=1}^{n} S^{-1}(\varepsilon a_{ij})e_j^{m_1} \quad \text{for } \eta + 1 \leq i \leq k.
\]

Hence, Equation 6.14 gives

\[
\|D_1(x_1)\|_w = \left\{ \sum_{i=1}^{n} \left[ S'(S^{-1}(\varepsilon)) \right]^{1/r} \right\}^{1/r} \quad (6.15)
\]

Next, if the components of the vectors \( y_i \) are denoted by \( y_{ij}, 1 \leq j \leq m_2 \), \( 1 \leq i \leq \eta \), and \( \tilde{Y} = \max_{i,j} |y_{ij}| \), then by Equations 4.4 and 6.5,

\[
\|W_2\|_{\infty} = \frac{1}{\varepsilon} \tilde{Y} \quad (6.16)
\]

From Inequalities 6.4 and Equation 4.3 one obtains

\[
\|W_1\|_{\infty} \leq \| ZX^{-1} \|_{\infty} \leq \| Z \|_{\infty} \| X^{-1}\|_{22}. \quad (6.17)
\]

By Equations 6.9 and 4.2, and the definition of \( Z \),

\[
\|Z\|_{\infty} = \max_{1 \leq i \leq \eta} \left[ \left( S^{-1}(\varepsilon) \right)^2 + \sum_{j=\eta+1}^{k} \left( S^{-1}(\varepsilon a_{ji}) \right)^2 \right]^{1/2} \quad (6.18)
\]
Finally, since $\|X^{-1}\|_2 = \frac{1}{\sigma_{\min}(X)}$ (Reference 8), by combining Equations 6.15 through 6.18 and Inequality 6.17 with Inequality 6.13 with $p = \infty$, $q = 2$, $r = 1$, and $t = \infty$, one obtains the following upper bound for $\|F'(x_i)\|_\infty$.

$$\|F'(x_i)\|_\infty \leq \left(\frac{1}{\varepsilon} \tilde{Y}\right) \cdot \frac{1}{\sigma_{\min}(X)} \cdot \max \left[\left[ S^{-1}(e)\right]^2 + \sum_{j=1}^{\eta} \left[ S^{-1}(ea_{ij})\right]^2 \right]^{1/2} \cdot \|D_1(x_i)\|_1. \quad (6.19)$$

Under certain conditions on the activation function $S$, one can obtain a simpler but more conservative upper bound by considering $\tilde{a} = \max_{ij} |a_{ij}|$ and $\tilde{a} = \min_{ij} |a_{ij}|$. Note that if $S$ is strictly increasing and symmetric about the origin, then $|S^{-1}(ea_{ij})| \leq S^{-1}(e\tilde{a})$ for all $i, j$. Moreover, if we assume that $S'$ is strictly decreasing on $(-\infty, 0)$, then $S'(S^{-1}(e\tilde{a}_{ij})) \leq S'(S^{-1}(e\tilde{a}))$ for all $i, j$. Under this condition on $S$ and $S'$, it follows that

$$\|F'(x_i)\|_\infty \leq \left(\frac{1}{\varepsilon} \tilde{Y}\right) \cdot \frac{1}{\sigma_{\min}(X)} \cdot \left[ \left[ S^{-1}(e)\right]^2 + (k - \eta) \left[ S^{-1}(e\tilde{a})\right]^2 \right]^{1/2} \cdot Q,$$

where $Q = \begin{cases} S'(S^{-1}(e)) & \text{for } 1 \leq i \leq \eta \\ \eta \cdot S'(S^{-1}(e\tilde{a})) & \text{for } \eta + 1 \leq i \leq k. \end{cases}$

If Technique 1 is used to interpolate through a set $\Omega$ with $k = m_1 + 1$ points, then $[W_2; b_2] = Y_{\alpha} X_{\alpha}^{-1}([W_1; b_1])$, where the multi-index $\alpha = (1, 2, \ldots, k)$. Note that $[W_2; b_2]$ is a function of $[W_1; b_1]$. Setting $p = \infty$, $r = 2$, $t = \infty$, and $q = 1$ in Inequality 6.13 and applying Inequalities 6.4 to $[W_2; b_2]$, one obtains

$$\|F'(x)\|_1 \leq \|Y_{\alpha}\|_2 \|X_{\alpha}^{-1}([W_1; b_1])\|_2 \|D_1(x)\|_2 \|W_1\|_1.$$

This inequality can be written more explicitly as

27
As we pointed out in Section 5, application of Technique 1 requires a selection of the first layer of weights $[W_1; \beta_1]$ before computing $[W_2; \beta_2]$. We gave no guideline on how to select $[W_1; \beta_1]$. Here we suggest that after an initial random choice of $[W_1; \beta_1]$, one proceeds to select new choices for $[W_1; \beta_1]$ using iterative minimization procedures to reduce the RHS of Inequality 6.20 at the points of interpolation $x_i (1 \leq i \leq k)$, thereby reducing the sensitivity of the transfer function $F$ with each iteration. The details of such a sensitivity-minimization algorithm are still under investigation.

7. SUMMARY

Finding the weight of a feedforward layered neural network so that the resulting transfer function maps a set of inputs to a desired set of outputs was described as an interpolation problem. It was shown how to define the weights so that the net interpolates through a set of $m_0 + 1$ points, where $m_0$ is the number of inputs and the net has one hidden layer.

It was also shown how to select the last layer of weights of a multilayered net so that the net interpolates through a set of $m_{L-1} + 1$ points, where $m_{L-1}$ is the number of neurons in the layer preceding the output layer. These two approaches (Techniques 1 and 2) provide a partial solution to the interpolation problem posed in Section 2. Moreover, both of the numbers $m_0 + 1$ and $m_1 + 1$ serve as lower bounds for the interpolation capacity of nets with one hidden layer, and the number $m_{L-1} + 1$ is a lower bound for the interpolation capacity of nets with $L$ layers of weights when $L > 2$.

The Jacobian matrix of the transfer function was computed. Its operator norm was used as a sensitivity measure of the transfer function to variations in its input. The induced $(p, q)$ matrix norms were introduced together with some of their properties in order to obtain computable upper bounds on the norm of the Jacobian matrix at the points of interpolation. The results suggest
that small weights are required for low sensitivity. It was also suggested that
the freedom that exists in the selection of the first \( L - 1 \) layers of weights,
when using Technique 1 for interpolation, can be exploited in order to
minimize the sensitivity of the transfer function at the points of interpolation.
The details of such a minimization algorithm is a topic for further research.

Another problem that is still under investigation is whether the first
\( L - 1 \) layers of weights can be selected (as well as the last layer of weights) so
that the net interpolates through more than \( m_{L-1} + 1 \) points. For example,
when \( L = 2 \) and \( m_0 > m_1 \), the net can realize more than \( m_1 + 1 \) points; namely,
\( m_0 + 1 \) using Technique 2.
8. REFERENCES


