NPS ARCHIVE
1998
PRITCHETT, W.

DUDLEY KNOX LIBRARY
NAVAL POSTGRADUATE SCHOOL
MONTEREY CA 93943-5101

DUDLEY KNOX LIBRARY
NAVAL POSTGRADUATE SCHOOL
MONTEREY CA 93943-5101

Copyright

by

William Christopher Pritchett

1998

Neural Networks for Classification

by

William Christopher Pritchett, B.S.

Thesis

Presented to the Faculty of the Graduate School

of The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Engineering

The University of Texas at Austin

May 1998

NPS Archive
1998
Pritchett, W.

DUDLEY KNOX LIBRARY
NAVAL POSTGRADUATE SCHOOL
MONTEREY CA 93943-5101

Neural Networks for Classification

Neural Networks for Classification

by

William Christopher Pritchett, M.S.E.

The University of Texas at Austin, 1998

SUPERVISOR: Irwin W. Sandberg

In many applications, ranging from character recognition to signal detection to automatic target identification, the problem of signal classification is of interest. Often, for example, a signal is known to belong to one of a family of sets $C_1, \ldots, C_n$ and the goal is to classify the signal according to the set to which it belongs. The main purpose of this thesis is to show that under certain conditions placed on the sets, the theory of uniform approximation can be applied to solve this problem. Specifically, if we assume that sets C_j are compact subsets of a normed linear space, several approaches using the Stone-Weierstrass theorem give us a specific structure for classification. This structure is a single hidden layer feedforward neural network. We then discuss the functions which comprise the elements of this neural network and give an example of an application.

Table of Contents

Chapter One: Signal Classification 1
Automatic Target Recognition 1
Pattern Recognition 2
Chapter Two: Neural Networks 5
Chapter Three: Background 11
Metric Spaces 11
Topological Spaces 14
Linear Spaces 18
Normed Linear Spaces 19
The Hahn-Banach Theorem and Separation in Linear Spaces 21
Chapter Four: The Stone-Weierstrass Theorem 26
Chapter Five: Neural Network Approximation of Continuous Maps 35
Chapter Six: Approximation and Classification 43
Polynomial Networks 46
Exponential Networks 46
Continuous Sigmoidal Networks 46
Squashing Function Networks 48
Radial Basis Function Networks 53
Chapter Seven: Applications 56
Classifier Example 56
Conclusions 57
Bibliography 59
Vita 62

List of Figures

Figure One: Nonlinear model of a neuron 9

Figure Two: A feed-forward neural network 10

Figure Three: A general structure for approximation 38

Figure Four: A structure for h 40

Figure Five: Representation of a classifying function 44

Figure Six: A classifying network 45

1. Signal Classification

Signal classification is, quite simply, the process of examining a signal and determining a class, or group, from which it came. Humans perform many instances of signal classification each day, often without even knowing it. For example, one might read a signature (the signal) carefully to determine the author (the class). This might be a process that would be extremely hard for a computer to perform.

There are numerous applications in military, civilian, and academic problems that require the use of the field of signal classification. It would be fruitless to attempt to compile an exhaustive list of applications, so we will state and develop a few problems here in which the theory of signal classification plays an important role in the solution.

Automatic Target Recognition

The field of automatic target recognition is extremely important, primarily in the area of the military. The main purpose of automatic target recognition is the use of computer processing to detect and recognize signatures in sensor data [1]. These targets are most often in a cluttered environment and frequently in hostile territory. They may include such things as aircraft, missiles, tanks, or warships. The clutter in their background may come from temperature or pressure disturbances, atmospheric variations, topographical objects, or even other targets.

There are typically two steps to an automatic target recognition problem:

detection and identification. Usually some relatively fast and coarse method is used to detect an object from background noise, and a slower more precise method is used to identify it. Typical features that are required to be extracted from the target when it is detected often include its position, its size and shape, and its speed.

In order to measure these quantities, an automatic target recognition system will possess sensors such as high resolution cameras and complex radar arrays. These sensors will obtain data and send it to the processing portion of the system. The system will then determine first whether a target even exists and then attempt to identify the target.

It is immediately very clear that the second portion of the problem (the identification) is basically a pure classification problem. Once it is determined that a tank is found, for example, it is important to be able to quickly determine whether the tank is friendly or hostile. An automatic recognition system thus frequently consists of several modules, one of which is the classifier.

Usually the classifier is designed with the assumption that each input, once found, belongs to only one of the classes. This assumption will become important later because it will allow us to make use of some well-known mathematical theorems in order to determine when classification may be possible.

Pattern Recognition

A second application of the theory of signal classification is in the field of pattern recognition. This is an extremely broad field, concerning a wide range

of problems of practical interest, including character recognition and speech identification.

One classical application is the reading of characters written either by hand or by machine. This application has a wide range of uses in government and commercial industry. For example, computers used by the post office are able to indentify machine-written letters on envelopes in order to sort them. Another important area deals with financial institutions. In these cases, the problem typically deals with classifying an input character into one of the thirty-six classes formed by the characters in the alphabet and the ten numerals. The area of printing is usually prescribed, so it is easy to locate and segment the characters. Some form of sampling is usually done, and then an algorithm determines the character.

There are also several problems in the field of speech recognition that rely heavily on classification theory. These problems include the following: speaker identification, speaker verification, and isolated word recognition [16]. In a speaker verification system, the number of classes relates to the number of different individuals that one wishes to recognize. In isolated word recognition, the number of classes will depend on the "vocabulary" of the system and may be as large as 10,000.

Many problems dealing with pattern recognition are found in the area of medicine as well. There are many applications that result in continuous functions, two-dimensional gray scale images, and time-varying images. These include results from electocardiograms, electroencephalograms, and X-ray im-

ages, to name a few. Cell analyzers classify blood cells in a population and determine cell type. Signal classification routines are of enormous importance in gathering fast information from these and other biological data.

These are just some of the many real-world applications in which signal classification plays a very important role. This makes it necessary to develop routines which are capable of performing well in signal processing problems. It is in this light that we consider the problem of determining a structure suitable for classification.

2. Neural Networks

It has long been recognized that the human brain functions in a completely different way from the modern digital computer. There has been a great interest in studying how the human brain works and in determining whether it is feasible to design a model capable of solving problems in a similar manner. Ramón and Cajál in 1911 introduced the concept of neurons as the basic elements of the brain [11]. It has been determined that neurons process information one hundred thousand to one million times slower than a basic silicon gate chip. The brain compensates for this slower speed by possessing in the neighborhood of 10 billion neurons and 60 billion synapses, or interconnections between the neurons [21]. As a result the brain is capable of performing many tasks at rates much greater than even the fastest computer. It is in an attempt to emulate this capability of the brain that the field of neural networks, or artificial neural networks, was born.

The history of neural networks dates back to the 1940's, when McCulloch and Pitts in 1943 proposed a computational model of an element resembling a neuron [3]. After some initial research, the idea faded until interest began to return in the 1980's. Since then, the field of neural networks has grown rapidly, with interest from researchers in a number of fields ranging from engineering to physics to psychology.

A neural network, essentially, is a structure that attempts to model the way the brain performs some task and then to perform that task in a similar manner. The structure may be electronically built or simulated in software, for

example. A neural network will contain a large number of individual cells, which model the neurons, and a number of interconnections between them, which model the synapses. Often the information passed through the interconnections will be multiplied by constants in order to achieve a certain task. This is known as weighting. Haykin gives a definition as adapted from Aleksander and Morton in 1990:

> A neural network is a massively parallel distributed processor that has a natural propensity for storing experimental knowledge and making it available for use. It resembles the brain in two respects:
>
> 1. Knowledge is acquired by the network through a learning process.
> 2. Interneuron connection strengths known as synaptic weights are used to store the knowledge.

The learning process mentioned here is often an attempt to modify the interconnection weights in order to accomplish the designated task. This attempt compares with the well-known field of adaptive filter theory, where filter weights are adapted over time until they approach a steady-state value.

There are many benefits that arise from neural networks' inherent structure. The following are some of them (see [11]).

1. Nonlinearity. The functions performed by the neurons are nonlinear; therefore the entire network, which is a weighted connection of these neu-

rons, will also be nonlinear. This helps in modeling typical applications, which are often nonlinear.

2. Input-output Mapping. One way in which the values for the weights used in the interconnections of the neural network are obtained is by a process called training. An example input is given, and weights are chosen so that the error between the actual output and some known desired output is minimized. This training procedure is repeated until the values of the weights reach a steady state (if possible). Thus the neural network learns by creating an input-output mapping.

3. Adaptivity. A neural network has the property of adapting its synaptic weights in order to match a change in the surrounding environment. When it is operating in one environment, it may be retrained to operate in another environment which has only minimal changes. Further, a neural network operating in a nonstationary environment is able to adapt its weights in real time.

4. Evidential Response. A neural network, when faced with a choice, is often able not only to select the right choice, but to give a confidence about the choice it made. For example, a neural network used for classification and given an input signal may output the class for that signal as well as how sure it is that that is actually the correct class.

5. Fault Tolerance. Since each of the many neurons in a neural network stores an important bit of information, the network's power is distributed

over each of these neurons. This allows the network in theory to continue operating even when one of the neurons fails, though with some degradation in performance. Neural networks are thus often marked by a gradual decay in performance instead of a single catastrophic failure.

6. Uniformality of Analysis and Design. Because all neural networks are similar in a structural sense and the same notation is used in the applications of neural networks to different problems, they are in a sense universal. This is seen in the following properties:

    - Neurons are common to all neural networks.
    - This commonality allows for the sharing of information between neural networks in different applications.
    - It is possible to build modular networks easily simply by integrating the different modules. In other words, parts of different networks (or even entire networks) may be used easily in conjunction with one another to create a new network.

As neurons are the building blocks of a neural network, their modeling is most important. The basic design for a neuron is fairly simple. A set of synapses are input to the neuron. These interconnections are weighted by real numbers, the synaptic weights. These weighted values are then summed. Finally, this sum is passed through a (typically) nonlinear activation function. This function usually serves to limit the output of the neuron to some desired range, for example $[0, 1]$ or $[-1, 1]$. An example of this model of a neuron is shown in

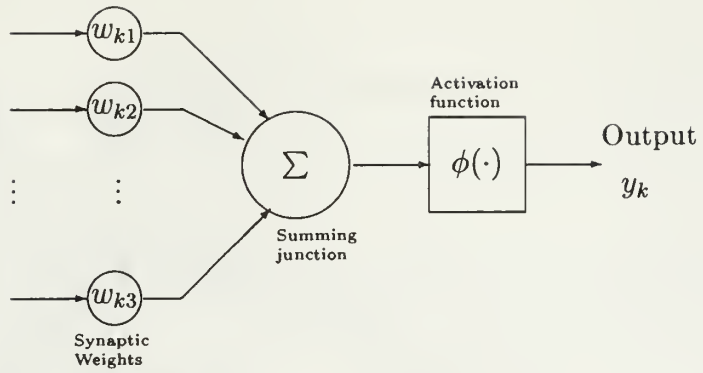

Figure 1: Nonlinear model of a neuron

Figure 1.

While the neurons themselves are modeled more or less the same regardless of the application, there are different architectures for the actual network. We will be concerned with just one particular type, called a feed-forward network with one hidden layer. This network architecture consists of a large number of neurons arranged schematically in three layers. This may be seen in Figure 2.

In theory, each unit of the input layer may be connected to each unit of the hidden layer. This connection has a weight, which as mentioned above is a real number, associated with it. The weights are denoted by w_{ij}. So each unit on the hidden layer receives a weighted sum of elements from the input layer and then processes this sum with an activation function. Finally, the result of this activation is transmitted to the output layer with another set of weights and then summed. The result for the network structure shown in Figure 2 is:

$$\sum_{i=1}^{n} \theta_i \sigma(\sum_{j=1}^{s} w_{ij}\zeta_{ij}).$$

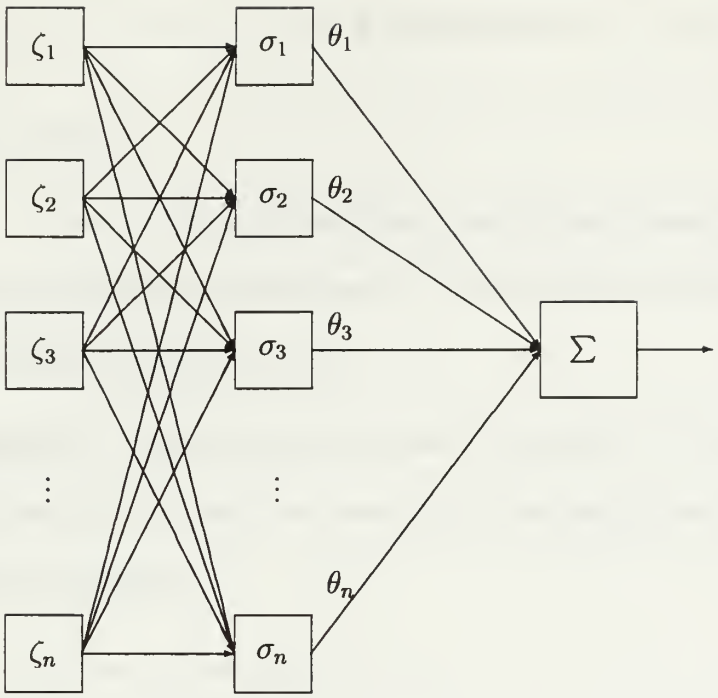

Figure 2: A feed-forward neural network

Finally, it is important to note that it is not necessarily possible to solve any problem simply by constructing a neural network at random and then attempting to train the weights. It is important to determine when a solution will be possible and what structure of network to try. Later it will be shown that a certain type of neural network is capable of solving an important classification problem.

3. Background

Metric Spaces

A type of space that will play a particularly important role in the study of approximation is a metric space. They are described in detail in many books, for example [9], [13], and [18].

Definition: A *metric space* is a pair (X, ρ) where X is a set of elements and ρ is a metric, or distance function, that is nonnegative and real-valued with the following properties:

1. $\rho(x, y) = 0$ if and only if $x = y$;
2. $\rho(x, y) = \rho(y, x)$;
3. $\rho(x, y) + \rho(y, z) \leq \rho(x, z)$.

Some examples of metric spaces are:

Example 1: The set of real numbers with metric $\rho(x, y) = |x - y|$, referred to as $\mathbb{R}$ or $\mathbb{R}^1$.

Example 2: The set of all ordered n-tuples $x = (x_1, x_2, \ldots x_n)$, with metric $\rho(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$. This space is generally referred to as $\mathbb{R}^n$.

Example 3: The set of continuous functions defined on a closed interval $[a, b]$ with metric $\rho(f, g) = \max_{a \leq t \leq b} |f(t) - g(t)|$.

Example 4: This same set of continuous functions along with the metric

$$\rho(f,g) = (\int_a^b [f(t) - g(t)]^2 \, dt)^{1/2}$$

form a different yet equally valid metric space (known as $L^2(\mathbb{R}^n)$. Thus, the metric as well as the set of points must be known in order for the space to be completely determined.

Let X be a metric space with $x_0 \in X$ and let $r > 0$. We define an *open ball* with radius r centered about x_0 (written $b(x_0, r)$) to be the set of points $x \in X$ such that $\rho(x, x_0) < r$. Let $A \subseteq X$. We define a point $x \in A$ to be an *interior point* of the set A if $b(x, r) \subseteq A$ for some $r > 0$. That is, we can find an open ball surrounding the point x such that every point in the ball belongs to the set A. It is in this way that we go about defining open sets in a metric space. In fact, a set $A \subseteq X$ is called an *open set* if all of its points are interior points.

Example 1: Consider the set $(0, 1)$ in $\mathbb{R}$. Given any point in the set, it is possible to choose an open ball of some radius such that the ball is contained in $(0, 1)$. Therefore, $(0, 1)$ is open in $\mathbb{R}$.

Example 2: On the other hand, consider the set $[0, 1)$ in $\mathbb{R}$ and look at any open ball about the point 0 with radius r. Whatever the choice of r, there will be points contained in the ball that are not in $[0, 1)$ (for example, the point $-r/2$); therefore the point 0 is not an interior point of the set $[0, 1)$, Therefore the set is not open.

Let X be a metric space and $x \in X$. We define a *neighborhood* of x as a

set containing an open set containing x. This open set will necessarily contain an open ball $b(x_0, \epsilon)$ for some $\epsilon > 0$. Therefore, every neighborhood of a point will contain an open ball of that point. Again let X be a metric space and let $A \subseteq X$. A point $x \in X$ is called a *contact point* of A if every neighborhood of x contains at least one point in A. Obviously all $x \in A$ are contact points of A. If every neighborhood of x contains infinitely many points in A, then x is called a *limit point* of A. Note that a limit point is necessarily a contact point by definition. The *closure* of a set A, written as $\overline{A}$, is simply the set of all the contact points of A. A set which is equivalent to its closure, $(A = \overline{A})$ is known as a *closed set.*

Example 1: Consider again the set $[0, 1)$ in $\mathbb{R}$. It is not possible to find an open ball about the point 1 that does not contain any points in $[0, 1)$. Therefore every neighborhood of 1 contains at least one point (in fact, every neighborhood contains infinitely many points) in the set $[0, 1)$. This implies that 1 is a contact point (and a limit point) of the set $[0, 1)$. Since $1 \notin [0, 1)$, the set does not coincide with its closure (in fact, as expected, $\overline{[0, 1)} = [0, 1]$) and is therefore not a closed set.

Example 2: On the other hand, the set $[0, 1]$ can be shown to be closed as its closure is the very same set $[0, 1]$.

One of the most important concepts concerning metric spaces is that of continuity. Let (X, ρ_x) and (Y, ρ_y) be metric spaces and let f be a function such that $f : X \to Y$. Then f is *continuous* at the point $p \in X$ if for every $\epsilon > 0$

there exists a $\delta > 0$ such that $\rho_y(f(x), f(p)) < \epsilon$ whenever $\rho_x(x, p) < \delta$.

A sequence $\{x_n\}$ in a metric space X is said to *converge* if there is a point $p \in X$ with the following property: For every $\epsilon > 0$ there is an integer N such that $n \geq N$ implies that $\rho(x_n, p) < \epsilon$. We write this as $x_n \to p$ or $\lim_{n\to\infty} x_n = p$. We define $\{x_n\}$ to be a *Cauchy sequence* in a metric space X if for every $\epsilon > 0$, there exists a positive integer N such that $|x_n - x_m| < \epsilon$ for $n, m > N$. We can easily show that a sequence converges if and only if it is a Cauchy sequence. A metric space is said to be *complete* if every Cauchy sequence converges to a point in the space. The completeness of certain metric spaces is very important to proving results in those spaces.

In a similar manner, we say that a sequence of functions $\{f_n\}$ from X to $\mathbb{R}$ converges uniformly on X to a function f if for every $\epsilon > 0$ there exists an integer N such that $n \geq N$ implies $|f_n(x) - f(x)| < \epsilon$ for all x. We often write this as $f_n \to f$ uniformly. For a discussion in greater depth of convergence, see [19].

Topological Spaces

Although metric spaces are usually the most general space needed, there may be times when a result may be proved for a more general space. It is for this purpose that we now introduce the topological space.

Definition: A *topological space* is the pair (X, τ) consisting of a set of points X and a topology τ, where τ is a family of subsets $G \subseteq X$, called open sets, with the following properties:

1. The set X itself and the empty set $\emptyset$ belong to τ.

2. Arbitrary unions $\bigcup_{\alpha} G_\alpha$ and finite intersections $\bigcap_{k=1}^{n} G_k$ of open sets belong to τ.

The definitions of open and closed sets in a topological space X is quite simple. A set $A \subseteq X$ is an *open set* if A belongs to τ. A set B in a topological space X is a *closed set* if its complement $X - B$ is open.

We can also extend the concepts of a neighborhood, contact point, limit point, and closure of a set in a topological space. By a *neighborhood* of x, we mean any open set G containing x. A point $x \in X$ is a *contact point* of $T \subseteq X$ if every neighborhood of x contains at least one point in T. A point $x \in X$ is a *limit point* of $T \subseteq X$ if every neighborhood of x contains infinitely many points in T. Finally, the *closure* of a subset T of a topological space X is the set of all the contact points of T.

Two important types of topological spaces are Hausdorff spaces and normal spaces. A topological space X is called a *Hausdorff space* if:

1. Sets consisting of single points are closed.

2. For every pair of distinct points x and y in X, there are disjoint neighborhoods of x and y.

A topological space is called a *normal space* if:

1. Sets consisting of single points are closed.

2. For every pair of disjoint closed sets A and B, there are disjoint neighborhoods of A and B.

Obviously, every normal set is Hausdorrf, though a Hausdorff set need not be normal. It can be verified that all metric spaces are topological spaces simply by taking τ to be the family of open sets that are open in the metric space in the usual sense. This is very important as it allows any result relating to topological spaces to be applied to metric spaces as well. In fact, we get an even better result: all metric spaces are normal (and therefore Hausdorff). The contrasts, however, to both of these statements are not true.

Example : The topological space consisting of only two points $\{0, 1\}$ where τ consists only of the sets $\{0, 1\}$ (the entire space) and $\emptyset$ is not a metric space.

Continuity in a topological space is a somewhat different concept than continuity in a metric space as well. Let (X, τ_x) and (Y, τ_y) be two topological spaces and let $f : X \to Y$. Then f is *continuous* if $f^{-1}(A) \in \tau_x$ for every A in τ_y. In other words, continuity implies that the inverse image of an open set is open.

A family $\mathcal{M}$ of subsets M_α of a topological space X is called a *cover* of X if $X \subseteq \bigcup_\alpha M_\alpha$. If the sets M_α consist entirely of open sets, then we call the family an *open cover*. A topological space is *compact* if every open cover has a finite subcover.

Although metric spaces possess many of the nice properties that we would like to have for topological spaces, it is not true that all metric spaces

are compact. There are some theorems (see for example [14]), however, that allow us to determine whether a given metric space is compact without having to view it as a topological space.

Let A and B be subsets of the metric space X. Then the set A is called an *ϵ-net* for the set B if there exists a point $x_a \in A$ such that for $\epsilon > 0$ any $x \in B$, $\rho(x, x_a) < \epsilon$.

Theorem 1 (Hausdorff). For compactness of a set M of a metric space X it is necessary that there should exist a finite ϵ-net of the set M for every $\epsilon > 0$. If the space X is complete, then the condition is also sufficient.

Roughly speaking, a set is compact if we can find a finite number of points and take open balls centered at those points such that the union of all the open balls contains the set. There are some improvements to this if we consider certain specific spaces.

Example 1: (Heine-Borel). A subset of $\mathbb{R}$ is compact if and only if it is closed and bounded.

Example 2: (Arzela). The functions of a set A are said to be *uniformly bounded* if there exists a constant K such that $|x(t)| < K$ for all $x(t) \in A$. The same functions are *equicontinuous* if given $\epsilon > 0$, there exists a $\delta > 0$ such that $|x(t_1) - x(t_2)| < \epsilon$ whenever $|t_1 - t_2| < \delta$. A set $A \subseteq C[0, 1]$, the space of real-valued continuous functions on the closed interval $[0, 1]$, is compact if and only if A is closed and the functions $x \in A$ are uniformly bounded and

equicontinuous.

Linear Spaces

We now introduce the concept of a linear space.

Definition : A nonempty set L is called a linear space if it satisfies the following axioms:

1. Any two elements $x \in L$, $y \in L$ uniquely determine a third element $x+y \in L$ called the sum of x and y that satisfies the following properties:

   (a) $x + y = y + x$ (commutativity);

   (b) $(x + y) + z = x + (y + z)$ (associativity);

   (c) L contains an element 0, called the zero element such that for all $x \in L$, $x + 0 = x$;

   (d) For each $x \in L$, there exists an element $-x \in L$ such that $x+(-x) = 0$, where 0 is the zero element;

2. There exists a product operation such that any element $x \in L$ and any number α determine a unique element $\alpha x \in L$ such that:

   (a) $\alpha(\beta x) = (\alpha\beta)x$

   (b) $1x \in L$;

3. The operations of addition and multiplication obey the following distributive axioms:

(a) $(\alpha + \beta)x = \alpha x + \beta x$;

(b) $\alpha(x + y) = \alpha x + \alpha y$.

The elements x, y, ... of a linear space are often called vectors, and the entire space is often called a vector space. The numbers α, β, ... are referred to as scalars and the entire set of allowable scalars is referred to as the field. Typically, the field is the set of real numbers, in which case the space is referred to as a real linear space. A subset L_0 of a linear space L is referred to as a linear subspace of L if L_0 itself is a linear space over the same field as L.

It is possible that a linear space possess no topology whatsoever as long as it satisfies the three properties above. However, in many applications the concepts of a linear space and topological space are combined. A space that is both a linear space and a topological space is referred to either as a linear topological space or a topological vector space. We require additionally only that the vector operations of addition and multiplication (which are not always the usual addition and multiplication) be continuous in the topology τ. It is possible too to apply the concept of a metric to a linear space, but what is more useful is to define an operation a bit more specific than a metric, called a norm, and apply it to a linear space.

Normed Linear Spaces

Definition: A linear space L equipped with an operation called a norm $(\|\cdot\|)$ is called a *normed linear space* if $\|\cdot\|$ satisfies the following three properties:

1. $\|x\| \geq 0$ for all x where $\|x\| = 0$ if and only if $x = 0$;

2. $\|\alpha x\| = |\alpha| \, \|x\|$ for all $x \in L$ and all α;

3. $\|x + y\| \leq \|x\| + \|y\|$ for all x and y in L.

Just as every metric space is also a topological space, every normed linear space may also be considered a metric space (and therefore a topological space as well) by taking the metric to be:

$$\rho(x, y) = \|x - y\|.$$

Again, the converse is not true.

Example: The metric space consisting of the closed interval $[0, 1]$ with the "discrete metric" $\rho(x, y) = 1$ if $x \neq y$ and $\rho(x, x) = 0$ cannot be made into a normed linear space.

A normed linear space that is complete (in the same sense that a metric space is complete) is known as a *Banach space.*

One special Banach space is called a Hilbert space.

Definition: A *Hilbert space* is a Banach space with the norm $\|x\| = < x, x >^{1/2}$ where $< \cdot, \cdot >$ is an inner product with the following properties (assuming the space is real):

1. $< x, y > \; = \; < y, x >$

2. $< \alpha_1 x_1 + \alpha_2 x_2, y > \; = \; \alpha_1 < x_1, y > + \alpha_2 < x_2, y >$

3. $< x, x > \ > 0$ for all $x \neq 0$.

The most common example of a Hilbert space is the n-dimensional space $\mathbb{R}^n$, with the Euclidean norm $\|x\| = \sqrt{\sum_{k=1}^{n} x_k^2}$ where $x = (x_1, x_2, \ldots, x_n)$.

The Hahn-Banach Theorem and Separation in Linear Spaces

One of the most important and fundamental results in all real analysis is the Hahn-Banach theorem. There are many different forms of the theorem and in most cases any version of the theorem can be used to directly prove any other version. It is first necessary to introduce the idea of convex sets and convex functionals.

Definition: A set $M \subset L$ is called a *convex set* if for each pair of points x, $y \in M$, all points on the line segment joining x and y (that is, all points of the form $kx + (1 + k)y$, $0 \leq k \leq 1$) are also elements of M.

Definition: A functional p defined on a real linear space L is said to be convex if it has the following properties:

1. $p(\alpha x) = \alpha p(x)$ for all $x \in L$ and all $\alpha \geq 0$;

2. $p(x + y) \geq p(x) + p(y)$ for all x, $y \in L$.

We now turn to the idea of extending a linear functional. Suppose we have a linear functional defined on a certain subspace. We want to know whether there exists a linear functional on the entire space that is equal to our first

functional on the subspace. The Hahn-Banach theorem tells us when this is possible.

Theorem 2 (Hahn-Banach) Let p be a finite convex functional defined on a real linear space L and let L_0 be any subspace of L. Let f_0 be any linear functional on L_0 satisfying the condition

$$f_0(x) \leq p(x)$$

on L_0. Then there exists a linear functional f on L, called the extension of f_0 such that $f = f_0$ at every point of L_0 and $f(x) \leq p(x)$ on L.

Proof: We can assume that $L_0 \neq L$. Let z be any element of $L - L_0$, and let $\tilde{L}$ be the subspace generated by L_0 and the element z, this being the set of all linear combinations of the form $x+tz$ ($x \in L_0, t \in \mathbb{R}$). For $\tilde{f}$ to be an extension of f_0 onto $\tilde{L}$, we need

$$\tilde{f}(x+tz) = \tilde{f}(x) + \tilde{f}(tz) = f_0(x) + t\tilde{f}(z)$$

Now, let $c = \tilde{f}(z)$ and note that if $\tilde{f}$ is an extension onto $\tilde{L}$ then $f_0(x) + tc \leq p(x+tz)$. This condition can easily translate to the two conditions:

$$c \leq p(x/t+z) - f_0(x/t) \text{ if } t > 0 \text{ and } c \geq -p(-x/t-z) - f_0(x/t) \text{ if } t < 0$$

So what remains is to show that there is always a c satisfying these conditions. In this light, let y_1 and y_2 be elements of L_0. Then

$$\begin{aligned} f_0(y_2 - y_1) &= f_0(y_2) - f_0(y_1) \leq p(y_2 - y_1) \\ &= p((y_2+z) - (y_1+z)) \leq p(y_2+z) + p(-y_1-z). \end{aligned}$$

So we get

$$-f_0(y_2) + p(y_2 + z) \geq -f_0(y_1) - p(-y_1 - z).$$

Now let $c_1 = \sup_{y_1}[-f_0(y_1) - p(-y_1 - z)]$ and $c_2 = \inf y_2[-f_0(y_2) + p(y_2 + z)]$. Then $c_2 \geq c_1$ and it simply remains to choose $c_2 \geq c \geq c_1$ and note that c satisfies the necessary conditions. So the functional f_t defined on L_t satisfies the condition $f(x) \leq p(x)$ for $x \in \tilde{L}$. An induction argument not given here proves the case when $\tilde{L}$ is the entire space L.

By applying the Hahn-Banach theorem, we may show a somewhat more useful result, given in [2].

Theorem 3 Let f be a bounded linear functional defined on the subspace L of the real normed linear space X. Then, there exists a bounded linear functional F defined on the entire space X so that $F(x) = f(x)$ for $x \in L$ and $\|F\| = \|f\|$.[1]

Proof: Since f is a bounded linear functional, then for $x \in L$, $|f(x)| \leq \|f\|\|x\|$. For $x \in X$ define $p(x) = \|f\|\|x\|$. It is then easy to show that p is convex and that $f(x) \leq p(x)$. By the Hahn-Banach Theorem, extend f to a new functional F defined on all of X such that $F(x) \leq p(x) = \|f\|\|x\|$ and $F(x) = f(x)$ for $x \in L$. Clearly, F is bounded and $\|F\| \leq \|f\|$. Similarly, if $x \in L$, then $|f(x)| = |F(x)| \leq \|F\|\|x\|$, implying $\|f\| \leq \|F\|$. Combining the two inequalities, we see that $\|F\| = \|f\|$ and the proof is complete.

[1]The norm operator $\|\cdot\|$, when applied to a bounded linear functional on a normed linear space X (as is the case here) is defined as $\|f\| = \sup_{\|x\| \leq 1} |f(x)|$. Further, $\|f\|$ can easily be shown to have the following properties: $\|f\| = \sup_{x \neq 0} \frac{|f(x)|}{\|x\|}$, and $|f(x)| \leq \|f\|\|x\|$ for all $x \in X$.

We now turn to perhaps the most useful corollary of the Hahn-Banach theorem. It is very desirable in many situations to know that there are a sufficient number of bounded linear functionals defined on a space to strictly separate the elements of that space. By strictly separate, we mean that for any two elements x_1 and x_2 of a linear space X, there exists an $f \in X^*$, the set of bounded linear functionals on X, such that $f(x_1) - f(x_2) \neq 0$. We prove this in the context of the following theorem.

Theorem 4 Let X be a normed linear space and $x_0 \in X$, $x_0 \neq 0$. Then there exists an $F \in X^*$ such that $\|F\| = 1$ and $F(x_0) = \|x_0\|$.

Proof: Let L be the linear subspace of X generated by taking the linear span of x_0. All elements in L will thus have a representation αx_0, $\alpha \in \mathbb{R}$. Define the function f on L by $f(\alpha x_0) = \alpha\|x_0\|$. It is seen at once that $f(x_0) = \|x_0\|$ simply by taking $\alpha = 1$. We can then extend f to a bounded linear functional F defined on the whole space X as noted in the previous theorem. Since $F{=}f$ on L, $F(x_0) = f(x_0) = \|x_0\|$. It thus remains only to show that $\|F\| = 1$. For any $x \in L$, we see that

$$|f(x)| = |f(\alpha x_0)| = |\alpha|\|x_0\| = \|\alpha x_0\| = \|x\|,$$

implying that $\|f\| = 1$ and therefore $\|F\| = 1$ by the previous theorem.

To prove our assertion about the strict separation of elements in a linear space by the functionals defined on that space, let X be a normed linear space and x_1 and x_2 be distinct elements in X. Further, let $f \in X^*$. Now define

$x_0 = x_1 - x_2$ and see that $x_0 \neq 0$ since x_1 and x_2 are distinct. We may now apply the previous theorem to get

$$f(x_1 - x_2) = f(x_0) = \|x_0\| \neq 0.$$

4. The Stone-Weierstrass Theorem and Uniform Approximation

In many applications, it is desirable to know whether a certain class of $\mathbb{R}$-valued functions may be useful in uniformly approximating a larger group of $\mathbb{R}$-valued functions. Weierstrass proved that it is possible to uniformly approximate any continuous functional on a compact subset of $\mathbb{R}^n$ by a polynomial in n variables. Since that time, there have been several different proofs of Weierstarass' theorem. One of the most useful is the one given by M. H. Stone in [23]. His primary result, which will be shown, generalizes Weierstrass' result in that it allows the domain to be any compact set (instead of just any compact subset of $\mathbb{R}^n$) and the set of approximating functions to be a set other than polynomials (which may not have meaning on a general compact set).

In order to generalize the theorem, we can view the polynomials as a subset of the set from which we obtain the approximating functional. We seek to know what functions may be derived from a certain set of prescribed functions by the specified algebraic operations of addition, multiplication, multiplication by real numbers and uniform passage to the limit. The set of prescribed functions for the polynomials, for example, consists of just two functions: $f_1(x) = 1$ and $f_2(x) = x$ defined on a bounded closed interval X of $\mathbb{R}$. From these two functions and the algebraic operations alone, the set of all polynomials may be formed. Weierstrass' theorem then tells us that the uniform passage to the limit of this set (the polynomials) is the set of all continuous functionals on X. Equivalently, the set of continuous functionals is the uniform closure of the set of

polynomials, or the continuous functions on X may be uniformly approximated by the set of polynomials.

In order to begin proving this generalized theorem, it is instructive to consider the case of a general topological space X where the specified algebraic operations are the lattice operations $\vee$ and $\wedge$ defined to be:

$$f \vee g = \max(f, g) \qquad \text{and} \qquad f \wedge g = \min(f, g)$$

These form the functions h and k defined as:

$$h(x) = \max(f(x), g(x)) \qquad \text{and} \qquad k(x) = \min(f(x), g(x))$$

for any $x \in X$. Let C be the set of all continuous real functions on X and C_0 be a prescribed subfamily of C. We want to obtain the family $U(C_0)$ of all functions which can be formed from the functions in C_0 by the application of the specified algebraic operations and uniform passage to the limit. In the case of the lattice operations, it is easily observed that $U(C_0)$ is a part of C closed under uniform passage to the limit, that is

$$U(C_0) \subset C, \qquad U(U(C_0)) = U(C_0).$$

The first property may be shown by observing that the mappings

$$x \longrightarrow \max(f(x), g(x)) \qquad \text{and} \qquad x \longrightarrow \min(f(x), g(x))$$

are continuous. This follows from the continuity of f and g (necessarily true since C_0 is a subfamily of C) and the continuity of the max and min mappings. Now since the uniform limit of continuous functions is also a continuous function,

clearly $U(C_0) \subseteq C$. To show that $U(U(C_0)) = U(C_0)$, we can form $U(C_0)$ in two steps. First, let $U_1(C_0)$ be the set containing all the functions obtained by applying the lattice operations alone to the functions in C_0. Then let $U_2(C_0)$ be the set consisting of the functions obtained from those in $U_1(C_0)$ by uniform passage to the limit. Clearly,

$$C_0 \subseteq U_1(C_0) \subseteq U_2(C_0) \subseteq U(C_0).$$

It remains to show that $U_2(C_0)$ is closed under the allowable operations, and therefore $U_2(C_0) = U(C_0)$. Let f be a function which is a uniform limit of functions f_n in $U_2(C_0)$. Then f must also be in $U_2(C_0)$ since given $\epsilon > 0$, there exists a function g_n in $U_1(C_0)$ such that $|f_n - g_n| < \epsilon/2$ since $U_2(C_0)$ is, by definition, the functions obtained by passing those in $U_1(C_0)$ to a uniform limit. Also, $|f - f_n| < \epsilon/2$ since our definition of f was a uniform limit of f_n. Therefore, $|f - g_n| < \epsilon$ and f is a uniform limit of functions g_n in $U_1(C_0)$ and therefore a member of $U_2(C_0)$ We must now show that whenever f and g are in $U_2(C_0)$, then so are $f \vee g$ and $f \wedge g$. This can be done by observing that if f and g are uniform limits of functions f_n and g_n in $U_1(C_0)$, then $f \vee g$ and $f \wedge g$ are uniform limits of $f_n \vee g_n$ and $f_n \wedge g_n$, respectively.

Theorem 5 Let X be a compact space, C the family of all continuous real functions on X, C_0 an arbitrary subfamily of C, and $U(C_0)$ the family of all functions (necessarily continuous) generated from C_0 by the lattice operations and uniform passage to the limit. Then a necessary and sufficient condition for a function f in C to be in $U(C_0)$ is that, whatever the points x, $y \in X$ and

whatever the positive number ϵ, there exists a function f_{xy} obtained by applying the lattice operations alone to C_0 and such that

$$|f(x) - f_{xy}(x)| < \epsilon \text{ and } |f(y) - f_{xy}(y)| < \epsilon.$$

Proof: The necessity is obvious. A proof of the sufficiency, which is not complicated, is given in [23]. There, Stone also notes the following corollary to the theorem.

Corollary 1: If C_0 has the property that, whatever the points $x, y \in X$, $x \neq y$ and whatever the real numbers α and β, there exists a function f_0 in C_0 for which $f_0(x) = \alpha$ and $f_0(y) = \beta$, then $U(C_0) = C$.

This tells us that the way in which a function f acts on pairs of points in X determines whether it can be approximated $U(C_0)$. This observation leads to the following theorem.

Theorem 6 Let X be a compact space, C the family of all continuous (necessarily bounded) real functions on X, C_0 an arbitrary subfamily of C and $U(C_0)$ the family of all functions (necessarily continuous) generated from C_0 by the linear lattice operations and uniform passage to the limit. Then a necessary and sufficient condition for a function f in C to be in $U(C_0)$ is that f satisfy every linear relation of the form $\alpha g(x) = \beta g(y)$, $\alpha\beta \geq 0$, which is satisfied by all functions in C_0. The linear relations associated with an arbitrary pair of points x, y in X must be equivalent to one of the following distinct types:

1. $g(x) = 0$ and $g(y) = 0$;

2. $g(x) = 0$ and $g(y)$ unrestricted, or vice versa;

3. $g(x) = g(y)$ without restriction on the common value;

4. $g(x) = \lambda g(y)$ or $g(y) = \lambda g(x)$ for a unique value λ, $0 < \lambda < 1$.

Corollary 1: In order that $U(C_0)$ contain a nonvanishing constant function, it is necessary and sufficient that the only linear relations of the form $\alpha g(x) = \beta g(y)$, $\alpha\beta > 0$, satisfied by every function on C_0 be those reducible to the form $g(x) = g(y)$.

Proof: It is obvious that when $U(C_0)$ contains a nonvanishing constant function then conditions (1), (2), and (4) can never be satisfied, so only (3) must be considered.

Corollary 2: In order that $U(C_0) = C$, it is sufficient that the functions in X_0 satisfy no linear relation of the form (1)–(4) of Theorem 1.

This is an important corollary because in practice it is easy to consider a set of functions with the property that all functions do not satisfy all of the relations (1)–(4).

Definition: A family of arbitrary functions on a domain X is said to be a separating family (for that domain) if, whenever X and y are distinct points of X, there is some function f in the family with distinct values $f(x)$, $f(y)$ at these points.

Corollary 3: If X is compact and if C_0 is a separating family for X and contains

a nonvanishing constant function, then $U(C_0) = C$.

Proof: Since C_0 contains a nonvanishing constant function, it may satisfy only condition (3) of Theorem 2. However, since C_0 is a separating family, there is a function $f \in C_0$ such that $f(x) \neq f(y)$ for x, y in X. So condition (3) is not satisfied by all functions in C_0. Therefore none of the conditions are satisfied by C_0 and therefore $U(C_0) = C$.

We now consider the case where $U(C_0)$ is built from the functions in $C_0 \subset C$ using the operations of addition, multiplication, multiplication by real numbers (the linear ring operations), and uniform passage to the limit. If f and g are uniform limits of the sequences f_n and g_n respectively, the product fg is not in general the uniform limit of the sequence $f_n g_n$. We therefore require that the set C consist of the bounded continuous functions on X. Of course, this is satisfied automatically when X is compact. This leads to the general theorem.

Theorem 7 Let X be a compact space, C the family of all continuous (necessarily bounded) functions on X, C_0 an arbitrary subfamily of C and $U(C_0)$ the family of all functions generated from C_0 by the linear ring operations and uniform passage to the limit. Then a necessary and sufficient condition for a function f in C to be in $U(C_0)$ is that f satisfy every linear operation of the form $g(x) = 0$ or $g(x) = g(y)$ which is satisfied by all functions in X_0.

Proof: As a lemma, one can show (see [23]) that if f is in $U(C_0)$ then so is $|f|$. This means that f is the uniform limit of functions in C_0 subject to the

linear ring operations. Using a well known representation of the min and max functions:

$$\max(a,b) = \frac{1}{2}(a+b+|a-b|)$$

$$\min(a,b) = \frac{1}{2}(a+b-|a-b|)$$

we can now see that whenever f and g are in $U(C_0)$ then $f \vee g$ and $f \wedge g$ are in $U(C_0)$ as well. So $U(C_0)$ is closed under the linear lattice operations as well as the linear ring operations and uniform passage to the limit. Therefore the results in Theorem 2 are applicable here. It remains to show that every function in $U(X_0)$ cannot satisfy linear relations of the form given in condition (4) of Theorem 2. Assume $g(x) = \lambda g(y)$ for every function g in $U(C_0)$ and every x, y in X, for $0 < \lambda < 1$. Then for every f in $U(C_0)$, f^2 is also in $U(C_0)$ and the relations $f^2(x) = \lambda f^2(y)$ and $\lambda f^2(y) = \lambda^2 f^2(y)$ would hold, implying that either $f(y) = 0$ for every f in $U(C_0)$ or $\lambda = 0, 1$, the second being a contradiction to the assumption. So we conclude that f is in $U(C_0)$ if and only if it satisfies all relations of the form $g(x) = 0$ or $g(x) = g(y)$ satisfied by those functions in C_0.

We give a definition in order to restate the general theorem.

Definition: A family $\mathcal{A}$ of real functions defined on a set X is said to be an algebra if (i) $f + g \in \mathcal{A}$, (ii) $fg \in \mathcal{A}$, and (iii) $cf \in \mathcal{A}$ for all $f \in \mathcal{A}$, $g \in \mathcal{A}$ and for all real constants c, that is, if $\mathcal{A}$ is closed under addition, multiplication, and multiplication by real numbers.

An equivalent form of the general theorem that is often used in practice

is stated in [19] as follows:

Theorem 8 Let $\mathcal{A}$ be an algebra of real continuous functions on a compact set K. If $\mathcal{A}$ separates points of K and if $\mathcal{A}$ does not vanish at any point in K, then any real continuous function on K may be approximated by an element of $\mathcal{A}$.

An argument in [4] extends the theorem to certain normed linear spaces that are not necessarily compact.

Theorem 9 Let X be a normed linear space (or, indeed, any Hausdorff topological space). If $\mathcal{A}$ is a subalgebra of $C(X)$, the continuous functions on X, that contains constants and separates the points of X, then $\mathcal{A}$ is dense in $C(X)$.

Proof: Let f be any element of $C(X)$. We must prove that each neighborhood of f contains an element of $\mathcal{A}$. Let K be a compact set in X and ϵ a positive number. By restricting f and all members of $\mathcal{A}$ to the compact set K, we can apply the classical version of the Stone-Weierstrass Theorem in $C(K)$. Its conclusion is that the set

$$\{g|K : g \in \mathcal{A}\}$$

is dense in C(K). Hence there is an element g in $\mathcal{A}$ such that $\|f - g\|_K < \epsilon$.

Now we give some examples from Stone's original article.

Theorem 10 Let X be an arbitrary bounded closed subset of n-dimensional Cartesian space, the coordinates of a general point being $x_1, \ldots, x_n$. Any continuous real function f defined on X can be uniformly approximated by polynomials in the variables $x_1, \ldots, x_n$. In case the origin $x = (0, \ldots, 0)$, the function

f can be uniformly approximated by polynomials vanishing at the origin if and only if f itself vanishes at the origin. Otherwise f can be uniformly approximated by such polynomials without qualification.

This is the classical approximation theorem proved by Weierstrass.

Theorem 11 Let f be an arbitrary continuous real function of the real variable θ, $0 \leq \theta \leq 2\pi$, subject to the periodicity condition $f(0) = f(2\pi)$. Then f can be uniformly approximated on its domain of definition by trigonometric polynomials of the form

$$p(\theta) = \frac{a_0}{2} + \sum_{n=1}^{N} (a_n \cos n\theta + b_n \sin n\theta).$$

Theorem 12 Any continuous real function f, which is defined on the interval $0 \leq x < \infty$ and vanishes at infinity in the sense that $\lim_{x \to \infty} f(x) = 0$, can be approximated by functions of the form $e^{-\alpha x} p(x)$ where $p(x)$ is a polynomial.

Theorem 13 Any continuous real function f which is defined on the interval $-\infty < x < +\infty$ and which vanishes at infinity in the sense that

$$\lim_{x \to -\infty} f(x) = \lim_{x \to +\infty} f(x) = 0$$

can be uniformly approximated by functions of the form $e^{-\alpha^2 x^2} p(x)$ where $p(x)$ is a polynomial.

Several of these examples will prove useful shortly.

5. Neural Network Approximation of Continuous Maps

We now will examine a structure that has been proven in useful for approximation. The structure will be based almost entirely on a proof in [20]. We assume that we have a normed linear space X and a subset C that is nonempty and compact. We let X^* represent the set of bounded linear functionals on X and Y represent a set of continuous maps which are dense in X^* on C in the usual sense. That is, for each $\phi \in X^*$ and for some $\epsilon > 0$, there exists a $y \in Y$ such that $|\phi(x) - y(x)| < \epsilon$ for $x \in C$. Further, for $k = 1, 2, 3, \ldots$ we let D_k be any family of continuous maps $h : I\!R^k \mapsto I\!R$ such that given a compact $E \subset I\!R^k$ and any continuous $g : E \mapsto I\!R$ as well as $\sigma > 0$ there exists an $h \in D_k$ such that $|g(x) - h(x)| < \sigma$ for $x \in E$. Let U be any set of continuous maps $U : I\!R \mapsto I\!R$ such that given $\sigma > 0$ and any bounded interval $(\beta_1, \beta_2) \subset I\!R$ there exists a finite number of elements $u_1, \ldots, u_l$ of U for which $|\exp(\beta) - \sum_j u_j(\beta)| < \sigma$ for $\beta \in (\beta_1, \beta_2)$.

Theorem 14 (Sandberg) Let $f : C \mapsto I\!R$. Then the following conditions are equivalent.

(*i*) f is continuous.

(*ii*) Given $\epsilon > 0$ there are a positive integer k, real numbers $c_1, \ldots, c_k$, elements $u_1, \ldots, u_k$ of U, and elements $y_1, \ldots, y_k$ of Y such that

$$|f(x) - \sum_j c_j u_j[y_j(x)]| < \epsilon$$

for $x \in C$.

(*iii*) Given $\epsilon > 0$ there are a positive integer k, elements $y_1, \ldots, y_k$ of Y, and an $h \in D_k$ such that

$$|f(x) - h[y_1(x), \ldots, y_k(x)]| < \epsilon$$

for $x \in C$.

Proof: First, assume condition (i) holds. Let V be the set of all functions $v : C \mapsto \mathbb{R}$ such that

$$v(x) = \sum \exp(\phi_j(x)),$$

in which the sum is finite and $a_j \in \mathbb{R}$ and $\phi_j \in X^*$. To see that V constitutes an algebra as defined above, observe that

$$\exp(\phi(x)) \exp(\psi(x)) = \exp(\phi(x) + \psi(x)) = \exp(\phi + \psi)(x).$$

Taking $\phi = 0$ we can see that V contains constants. Finally, we have demonstrated previously that the Hahn-Banach theorem guarantees that we can choose an x and y in C such that $\phi(x - y) \neq 0$. Therefore, $\exp(\phi(x)) \neq \exp(\phi(y))$, so V separates the points of C. We may now apply the Stone-Weierstrass theorem guaranteeing uniform approximation on compacta. In other words, for $\epsilon > 0$, there are a positive integer n, real numbers $d_1, \ldots, d_n$, and elements $z_1, \ldots, z_n$ of X^* such that

$$|f(x) - \sum_{j=1}^{n} d_j \exp(z_j(x))| < \epsilon/3$$

for $x \in C$.

Assume that $\sum_j |d_j| \neq 0$. Choose $\gamma > 0$ such that $\gamma \sum_j |d_j| < \epsilon/3$. Let $[a', b']$ be an interval in $\mathbb{R}$ that contains all of the sets $z_j(C)$, and let $a \in \mathbb{R}$

and $b \in I\!R$ such that $a < a'$ and $b > b'$. That is, the interval $[a, b]$ contains the interval $[a', b']$. Now, choose $\nu > 0$ such that $|\exp(\beta_1) - \exp(\beta_2)| < \gamma$ for β_1, $\beta_2 \in [a, b]$ with $|\beta_1 - \beta_2| < \nu$. Clearly this is possible because of the continuity of the exponential function. Set $\rho = \min(\nu, a' - a, b - b')$ and choose $y_j \in Y$ such that $|z_j(x) - y_j(x)| < \rho$, $x \in C$ for all j. This gives $|\exp(z_j(x)) - \exp(z_j(x))| < \gamma$, $x \in C$ for each j. Now using a version of the triangle inequality, this gives:

$$\begin{aligned} |f(x) - \sum_j \exp(y_j(x))| &\leq |f(x) - \sum_j \exp(z_j(x))| \\ &+ |\sum_j \exp(z_j(x)) - \sum_j \exp(y_j(x))| \\ &\leq \epsilon/3 + \sum_j |d_j| |\exp(z_j(x)) - \exp(y_j(x))| \\ &\leq 2\epsilon/3, \end{aligned}$$

for $x \in C$.

Now we choose $u_1, \ldots, u_l \in U$ so that

$$|\exp(\beta) - \sum_i u_i(\beta)| \leq \gamma_1, \beta \in [a, b]$$

where $\gamma_1 \sum_j |d_j| < \epsilon/3$. Then,

$$|f(x) - \sum_j \sum_i d_j u_i[y_j(x)]| \leq |f(x) - \sum_j d_j \exp[y_j(x)]| + |\sum_j d_j \exp[y_j(x)]$$

$$- \sum_j \sum_i d_j u_i[y_j(x)]| \leq (2\epsilon)/3 + \sum_j |d_j \exp[y_j(x)] - d_j \sum_i u_i[y_i(x)]|$$

$$\leq (2\epsilon)/3 + \sum_j |d_j| |\exp[y_j(x)] - \sum_i u_i[y_i(x)]| \leq (2\epsilon)/3 + \gamma_1 \sum_j |d_j| < \epsilon.$$

Now, since $\sum_j \sum_i d_j u_i[y_j(x)]$ is equivalent to $\sum_j c_j u_j[y_j(x)]$, with the c_j, u_j, and y_j in $I\!R$, U, and Y, respectively, we have shown that $(i) \rightarrow (ii)$.

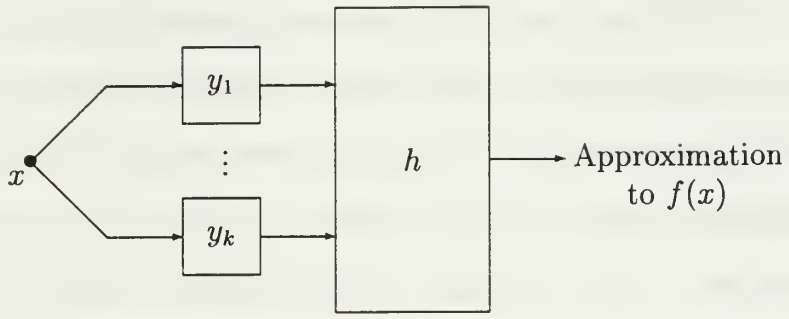

Figure 3: A general structure for approximation

To show that $(ii) \rightarrow (iii)$, let $\epsilon > 0$ and suppose that there exist k, $c_1, \ldots, c_k$, and $u_1, \ldots, u_k$ such that

$$|f(x) - \sum_j c_j u_j[y_j(x)]| < \epsilon/2, x \in C.$$

Let $h \in D_k$ satisfy $|h(\lambda) - \sum_j c_j u_j(\lambda)| < \epsilon/2$ for $\lambda \in [a,b]^k$. Then

$$|f(x) - h([y_1(x), \ldots, y_k(x)]| \leq |f(x) - \sum_j c_j u_j[y_j(x)]| +$$

$$|\sum_j c_j u_j[y_j(x)] - h(y_1(x), \ldots, y_k(x)]| < \epsilon/2 + \epsilon/2 = \epsilon$$

for $x \in C$.

Finally, $(iii) \rightarrow (i)$ as f is a uniform limit of continuous functions and therefore continuous itself.

This proof has demonstrated a general structure that may be used for approximation. This structure is shown below in Figure 3.

Part (iii) of the theorem shows that the y_j's are simply functions which are capable of approximating linear functionals defined on the space X (these

may actually be linear functionals themselves) while the structure for h is simply a continuous memoryless nonlinear system capable of approximating uniformly on compacta in $\mathbb{R}^k$. In other words, the problem of approximating a function whose domain may be any compact subset of any normed linear space has been reduced to the problem of approximating a function on $\mathbb{R}^k$, a subject about which a great deal is known, and has been shown to some extent in dealing with the Stone-Weierstrass theorem. Stiles, Sandberg, and Ghosh have shown in [22] that structures of a similar form have use in the approximation of certain nonlinear discrete time mappings as well.

Part (ii) of the theorem gives a specific example of the structure of the network. Again it takes the y_j's to be uniform approximations of linear functionals on X. Here one possible structure for h is shown as below in Figure 4. The u_j's, as mentioned before, are drawn from a set capable of uniform approximation of the exponential function on a bounded set in $\mathbb{R}$. In the simplest case, from the perspective of the theorem, each u_j may be taken to be the function $\exp(\cdot)$.

In a moment we will determine possible choices for the elements u_j in the approximation network. Now we will look at a similar method of dealing with this problem given in [4], [7], and [24]. We start be defining a certain class of functions, called ridge functions and then immediately give the theorem.

Definition: A function $f : X \mapsto \mathbb{R}$ is called a *ridge function* if it may be represented in the form $f = g \circ \phi$, where $g : \mathbb{R} \mapsto \mathbb{R}$ and $\phi \in X^*$, where X^* is the space of continuous linear functionals on X. An alternative equivalent form

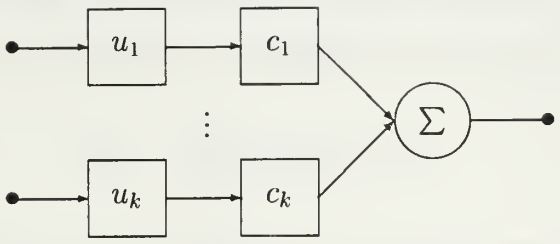

Figure 4: A structure for h

of this composite function is $f(x) = g(\phi(x))$ for $x \in X$.

It can easily be shown, for example, that all ridge functons on $\mathbb{R}^n$ can be written in the form

$$f(x) = g(a_1\zeta_1 + a_2\zeta_2 + \cdots + a_n\zeta_n)$$

where $x = (\zeta_1, \zeta_2, \ldots, \zeta_n) \in \mathbb{R}^n$.

Theorem 15 (Cheney) Let G be a fundamental set in $C(\mathbb{R})$[2] and let X be a normed linear space. Let Φ be a subset of X^* such that the set

$$\phi/\|\phi\| : \phi \in \Phi, \phi \neq 0$$

is dense in the unit sphere of X^*. Then the set of ridge functions $\{g \circ \phi : g \in G, \phi \in \Phi^*\}$ is fundamental in $C(X)$.[3]

[2]A subset Y of X is said to be fundamental in X if its linear span is dense in X. Thus, there are elements $y_1, \ldots, y_n \in Y$ such that for any $x \in X$ and $\epsilon > 0$, $|x - \sum_{j=1}^{n} c_j y_j| < \epsilon$ where $c_j \in \mathbb{R}$.

[3]$C(X)$ is, of course, the set of continuous, real-valued functions on the normed linear space X.

Proof: Let f be a member of $C(X)$, C a compact set in X, and $\epsilon > 0$. We have shown above that there exist $u_j \in C(\mathbb{R})$ and $y_j \in X^*$ such that

$$\|f(x) - \sum_{j=i}^{P} u_j \circ y_j\| < \epsilon/3$$

for $x \in C$. By adjusting the functions u_j as necessary, we can assume that $\|y_j\| = 1$ for $1 \leq j \leq m$. Let $M = \sup_{x \in C} \|x\|$. Choose $\delta > 0$ so that when $|s| \leq M$, $|t| \leq M$, and $|s - t| < \delta$ we get $|u_j(s) - u_j(t)| < \epsilon/3P$ for $1 \leq j \leq P$. This is, of course, possible because the u_j are continuous. Now select $\phi_j \in \Phi$ so that $\|\phi_j/\|\phi_j\| - y_j\| < \delta/M$ for $1 \leq j \leq P$. Let $\lambda_j = 1/\|\phi_j\|$ and $\mu = \max_j \|\phi_j\|$. Select $a_{jk} \in \mathbb{R}$ and $g_{jk} \in G$ so that for $|T| \leq \mu M$ we have

$$|u_j(\lambda_j t) - \sum_{k=1}^{N} a_{jk} g_{jk}(t)| < \epsilon/3P \qquad (1 \leq j \leq P).$$

Now let $x \in C$. Then $\|x\| \leq M$, $|y_j(x)| \leq M$, $|\lambda_j \phi_j(x)| \leq M$, and

$$|y_j(x) - \lambda_j \phi_j(x)| \leq \|x\| \|y_j\| \|y_j - \lambda_j \phi_j\| \leq M(\delta/M) = \delta.$$

From the definition of δ (i.e., let $s = y_j(x)$ and $t = \lambda_j \phi_j(x)$) we get

$$|\sum_{j=1}^{P} h_j(y_j(x)) - \sum_{j=1}^{P} h_j(\lambda_j \phi_j(x))| \leq \sum_{j=1}^{P} \epsilon/3P = \epsilon/3.$$

Now, because $|\phi_j(x)| \leq \|\phi_j\| \|x\| \leq \mu M$, the definition of a_{jk} and g_{jk} gives

$$|\sum_{j=1}^{P} h_j(\lambda_j \phi_j(x)) - \sum_{j=1}^{P} \sum_{k=1}^{N} a_{jk} g_{jk}(\phi_j(x))| \leq \sum_{j=1}^{P} \epsilon/3P = \epsilon/3.$$

Now, by a simple application of the triangle inequality, we get

$$|f(x) - \sum_{j=1}^{P} \sum_{k=1}^{N} a_{jk} g_{jk}(\phi(x))| \leq |f(x) - \sum_{j=1}^{P} h_j(y_j(x))|$$

$$+|\sum_{j=1}^{P} h_j(\lambda_j\phi_j(x)) - \sum_{j=1}^{P} h_j(\lambda_j(x))| + |\sum_{j=1}^{P} h_j(\lambda_j\phi_j(x)) - \sum_{j=1}^{P}\sum_{k=1}^{N} a_{jk}g_{jk}(\phi_j(x))| < \epsilon.$$

Since $\sum_{j=1}^{P}\sum_{k=1}^{N} a_{jk}g_{jk}(\phi_j(x))$ may be written as $\sum_{j=1}^{S} c_j g_j(\phi_j(x))$, we get the desired result:

$$|f(x) - \sum_{j=1}^{S} c_j g_j(\phi_j(x))| < \epsilon \text{ for } x \in C.$$

We note many similarities between this proof and part of Sandberg's. The set of functions G in Cheney's theorem is similar to the set of functions U in Sandberg's, but the requirement in Sandberg's theorem on U is less stringent. The set U is required only to approximate one specific function in $C(\mathbb{R})$, namely the exponential function, $\exp(\beta)$, on a certain bounded set. Cheney's theorem, on the other hand, requires that the set G be fundamental in $C(\mathbb{R})$. This means that any continuous function defined on a compact set in $\mathbb{R}$ is capable of being approximated by the set G.

6. Approximation and Classification

As previously mentioned, the problem of classifying signals plays an important role in a variety of problems. We attempt to provide the framework for a solution to some of these problems by restating the problem in a more mathematical sense.

We assume first that all of the signals to be classified are drawn from a normed linear space. For simplicity, we will further assume that each signal may belong only to one of the classes. For example, assume that there are n different classes $C_1, \ldots, C_n$ that are all subsets of a normed linear space X, and that each signal received must necessarily belong to exactly one of the classes.

We now have the framework whereby we can view the classifier as a mathematical function f that takes the signal to be classified as input and produces the desired class as output. For example, if $x \in C_j$, then $f(x) = a_j$, where $a_1, \ldots, a_n$ are all distinct integers, would model a classification system whereby each element of class C_j be mapped to the integer a_j. A graph of this simple function is shown in Figure 5. Our assumption that each signal may belong to only one class means that the sets C_j are pairwise disjoint.

In order to apply the theorems that we have developed, it is helpful to assume that the sets C_j are compact. This assumption will, of course, exclude certain classification problems from the scope of these theorems. We now can let $C = \bigcup_{j=i}^{n} C_j$. The set C will now also be compact as it is the union of a finite number of compact disjoint sets. Finally, since the function f is constant on each set C_j and the distance between any pair of sets is positive, the function

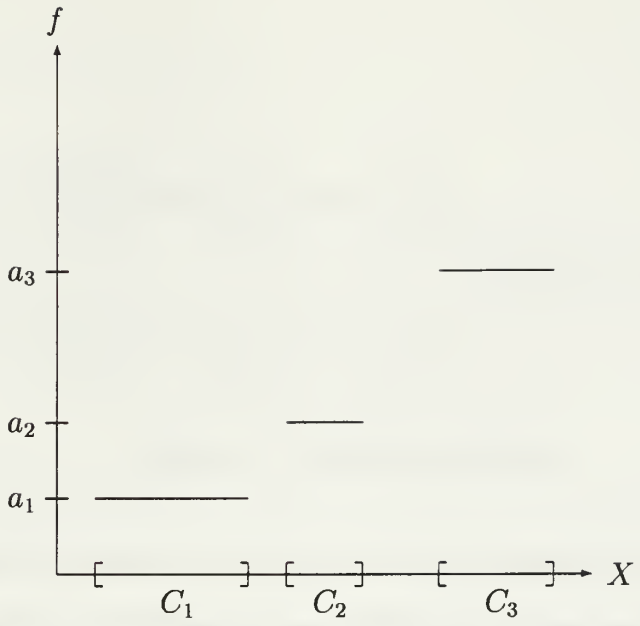

Figure 5: Representation of a classifying function

f is continuous. With these assumptions, we get the following:

1. There are real numbers $c_1, \ldots, c_k$, elements $y_1, \ldots, y_k \in Y$, a positive integer n, elements $u_1, \ldots, u_n$ of U and $\epsilon > 0$ such that

$$a_j - \epsilon < \sum_{j=1}^{k} c_j u_j[y_j(x)] < a_j + \epsilon$$

for $x \in C_j$ and $j = 1, \ldots, m$.

2. There are a positive integer k, elements $y_1, \ldots, y_k$ of Y and an $h \in D_k$ such that

$$a_j - \epsilon < h[y_1(x), \ldots, y_k(x)] < a_j + \epsilon$$

for $x \in C_j$ and $j = 1, \ldots, m$.

These follow directly from Sandberg's theorem.

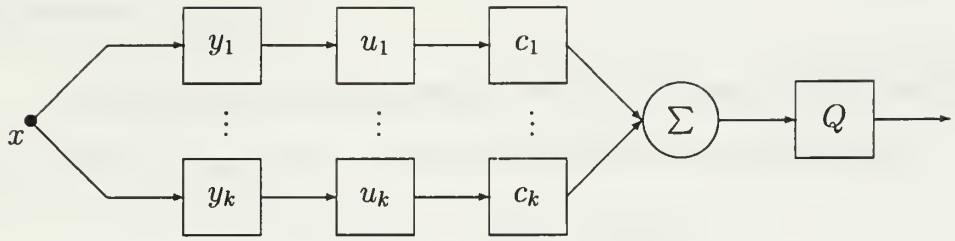

Figure 6: A classifying network

This now allows us to use the above approximation network for the purpose of classification. We require one additional element and that is a quantizer Q. This quantizer is simply a real functional $Q : \mathbb{R} \mapsto \mathbb{R}$ such that Q maps numbers in the interval $(a_j - \epsilon, a_j + \epsilon)$ to a_j. As long as we choose $\epsilon < 0.5 \min_{i \neq j} |a_i - a_j|$, then this quantizer, when following a network of the structure defined above, will allow the correct class to be output. This gives an entire structure for a classification network. It is shown in Figure 6. The structure for h as defined in part (ii) of Sandberg's theorem is used in the figure.

We now turn to demonstrating some acceptable choices for the hidden elements in our classification network. In all cases, the complete structure of the network is as in Figure 6. No assumption is made about the number n (how many elements are necessary) or the determination of the constants c_j. We are concerned entirely with determining suitable choices for the u_j and give several examples as well as a justification for each here. In each case, the y_j will be assumed to be either bounded linear functionals on X or elements capable of

uniformly approximating them.

Polynomial Networks

A polynomial network is simply one in which each u_j is a polynomial. In the ridge function form, a polynomial network will be of the form

$$\sum_i p_i \circ \phi_i = \sum_i \sum_j c_{ij}[\phi_j(x)]^j.$$

The original Weierstrass Approximation theorem showed that polynomials were capable of approximating on $\mathbb{R}$. Now, either Theorem 14 or Theorem 15 tells us that polynomials, when placed in the network, are capable of solving the classification problem.

Exponential Networks

An exponential network in which each of the elements u_j is of the form $\exp(\cdot)$ is the most basic to justify as the proof of Sandberg's theorem is based on showing first how the exponential functional is capable of being used as the nonlinear element and then showing how a function capable of uniformly approximating it on a bounded interval is also acceptable.

Continuous Sigmoidal Networks

A more complicated but extremely important type of network that is useful for classification is a continuous sigmoidal network. It is first necessary to define a sigmoidal function.

Definition : A functional $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is called a *sigmoidal function* or *sigmoid*

if

$$\lim_{t\to-\infty}\sigma(t)=0 \text{ and } \lim_{t\to\infty}\sigma(t)=1.$$

In 1989, Cybenko (see [8]) proved that for any compact set $C\subset I\!R^n$, any $f\in C(C)$, and for any $\epsilon>0$ there exists a function g of the form

$$g(x)=\sum_{j=1}^{m}a_j\sigma(<\gamma_j,x>+\theta_j) \qquad (x,\gamma_j\in I\!R^n,\theta_j\in I\!R)$$

where σ is a continuous sigmoidal function such that

$$|g(x)-f(x)|<\epsilon \text{ for all } x\in C.$$

In other words, this sum of translations and dilations of a sigmoidal function is capable of uniformly approximating any bounded continuous functional on a compact subset of $I\!R^n$. Sandberg mentions in [20] that given that the statement is true for $n=1$, the $(i)\to(ii)$ section of his proof quickly extends the result for $n>1$. Indeed, if we let X be simply $I\!R^n$, the elements y_j be linear functionals defined on $I\!R^n$, and $u_j(x)=c_j\sigma(\alpha_j x+\beta_j)$ where $c_j,\alpha_j,\beta_j\in I\!R$. This gives us a sum of the type desired for $n>1$.

In [5], Cheney demonstrates as a result of the general theory of ridge functions that the result is applicable when the elements of the vectors γ_j and the numbers θ_j are integers. In fact, the theorem is given as follows.

Theorem 16 Let g be a continuous function on $I\!R$ such that the limits of $g(t)$ as $t\to\infty$ and $t\to-\infty$ exist and are different. Put $g_{ij}=g(jt+i)$. Then $\{g_{ij}:i,j\in\mathcal{Z}\}$ is fundamental in $C(I\!R)$.

The proof of this theorem relies on measure theory, making use of the Riesz Representation Theorem and the Dominated Convergence Theorem. It is beyond the scope of this thesis but can be found in [4].

It is seen that this theorem allows g to be a continuous sigmoid, but does not require it. The only importance when using the translations and dilations is that the limits at ∞ and at $-\infty$ are not the same. It was mentioned earlier that often times it is desired that the output of the activation function in a neural network be in a certain range such as $[0, 1]$. Sigmoidal functions fit nicely into this framework.

Finally, we can show at once that these shifted and scaled sigmoidal functions are capable of approximating on any normed linear space by using either of the two main theorems after noting that they are capable of approximating on $\mathbb{R}$.

Squashing Function Networks

The previous section has dealt with the use of translations and dilations of continuous sigmoidal functions. In this section, we will deal with certain type of sigmoid that is not necessarily continuous, a squashing function, and attempt to obtain a similar result. A squashing function is defined in [12] as follows:

Definition: A function $\Psi : \mathbb{R} \mapsto [0, 1]$ is a *squashing function* if it is nondecreasing, $\lim_{\lambda \to \infty} \Psi(\lambda) = 1$, and $\lim_{\lambda \to -\infty} \Psi(\lambda) = 0$.

It is seen at once that this definition simply requires that Ψ be a nondecreasing sigmoidal function (not necessarily continuous). Some useful squashing

functions include the threshold function, $\Psi(\lambda) = 1_{\{\lambda \geq 0\}}$ where $1_{\{\cdot\}}$ is the indicator function; the ramp function, $\Psi(\lambda) = \lambda 1_{\{0 \leq \lambda \leq 1\}} + 1_{\{\lambda > 1\}}$; and the cosine squasher (see [10]), $\Psi(\lambda) = (1 + \cos[\lambda + 3\pi/2])(1/2)1_{\{-pi/2 \leq \lambda \leq \pi/2\}} + 1_{\{\lambda > \pi/2\}}$.

Hornik et al. first define what they call a sigma-pi network and prove certain results pertaining to it. Following this, they extend the results to a network resembling those that have been mentioned above. We proceed as did he, considering only the $\mathbb{R}^1$ case.

Definition: For any measurable function G mapping $\mathbb{R}$ to $\mathbb{R}$, let $\sum \prod^1(G)$ be the class of functions

$$\{f : \mathbb{R} \mapsto \mathbb{R} : f(x) = \sum_{j=1}^{q} \beta_j \prod_{k=1}^{l_j} G(A_{jk}(x)), \quad x, \beta_j \in \mathbb{R}, A_{jk} \in A, q = 1, 2, \ldots\}.$$

where $l_j \in \mathbb{N}$ and A is the set of all affine functions from $\mathbb{R}$ to $\mathbb{R}$, that is, the set of all functions of the form $A(x) = wx + b$ where w, $b \in \mathbb{R}$. Networks of this form are referred to as sigma-pi networks.

Definition: For any measurable function G mapping $\mathbb{R}$ to $\mathbb{R}$, let $\sum^1(G)$ be the class of functions

$$\{f : \mathbb{R} \mapsto \mathbb{R} : f(x) = \sum_{j=1}^{q} \beta_j G(A_j(x)), \quad x, \beta_j \in \mathbb{R}, A_j \in A, q = 1, 2, \ldots\}.$$

This form of this second network clearly resembles the continuous sigmoidal network that was shown above if G is taken to be a continuous sigmoidal function. The shifting and scaling that was present above is simply performed

by the affine functional here; only the notation is different. For now, we will continue to let G be any function.

We now give the main result that applies here.

Theorem 17 For every squashing function Ψ, $\sum^1(\Psi)$ is uniformly dense on compacta in $C(I\!R)$.

Proof: We proceed by first proving several lemmas that will aid in the proof.

Lemma 1: Let $G : I\!R \mapsto I\!R$ be continuous and nonconstant. Then $\sum\prod^1(G)$ is uniformly dense on compacta in $C(I\!R)$.

Proof: We can apply the Stone-Weierstrass Theorem here. Let $C \subset I\!R$ be any compact set. For any G, $\sum\prod^1(G)$ is obviously an algebra on C. If x, $y \in C$, $x \neq y$, then we can find an $A_1 \in A$ such that $G(A_1(x)) \neq G(A_1(y))$. To show this, pick a, $b \in I\!R$, $a \neq b$ such that $G(a) \neq G(b)$. Then choose $A_1(\cdot)$ to satisfy $A_1(x) = a$ and $A_1(y) = b$. Then $G(A_1(x)) \neq G(A_1(y))$. This ensures that $\sum\prod(G)$ is separating. Now we must show that $\sum\prod^1(G)$ vanishes on no point of C. Pick $b \in I\!R$ such that $G(b) \neq 0$ and $A_2(x) = 0 \cdot x + b$. For all $x \in C$, $G(A_2(x)) = G(b) \neq 0$, so this is a nonvanishing constant function. The Stone-Weierstrass theorem now guarantees that $\sum\prod^1(G)$ is capable of uniformly approximating any continuous functional on C.

This lemma shows that the sigma-pi networks are capable of uniform approximation of any continuous function on a compact set regardless of the G

with the only requirements that G be continuous and nonconstant. We have not yet required that G be a squashing function.

Lemma 2: Let F be a continuous squashing function and Ψ be an arbitrary squashing function. For every $\epsilon > 0$ there is an element H_ϵ of $\sum^1(\Psi)$ such that $\sup_{\lambda \in \mathbb{R}} |F(\lambda) - H_\epsilon(\lambda)| < \epsilon$.

Proof: Choose $\epsilon > 0$ and assume without loss of generality that $\epsilon < 1$. We must now find constants β_j and affine functions A_j, $j \in \{1, 2, \ldots, Q-1\}$ such that

$$\sup_{\lambda \in \mathbb{R}} |F(\lambda) - \sum_{j=1}^{Q-1} \beta_j \Psi(A_j(\lambda))| < \epsilon.$$

Choose Q such that $1/Q < \epsilon/2$. For $j \in \{1, 2, \ldots, Q-1\}$, set $\beta_j = 1/Q$. Pick $M > 0$ such that $\Psi(-M) < \epsilon/2Q$ and $\Psi(M) > 1 - \epsilon/2Q$. Such an M can be found because Ψ is a squashing function. For $j \in \{1, 2, \ldots, Q-1\}$, set $r_j = \sup\{\lambda : F(\lambda) = j/Q\}$. Set $r_Q = \sup\{\lambda : F(\lambda) = 1 - 1/2Q\}$. Because F is a continuous squashing function, such r_j's exist. Now, for any $r < s$, let $A_{r,s} \in A$ be the unique affine function satisfying $A_{r,s}(r) = M$ and $A_{r,s}(s) = -M$. The desired approximation is then $H_\epsilon = \sum_{j=1}^{Q-1} \beta_j \Psi(A_{r_j, r_{j+1}}(\lambda))$. We can easily check that on the intervals $(-\infty, r_1], (r_1, r_2], \ldots, (r_{Q-1}, r_Q], (r_Q, \infty)$, $|F(\lambda) - H_\epsilon(\lambda)| < \epsilon$.

Lemma 3: For every squashing function Ψ, every $\epsilon > 0$, and every $M > 0$ there is a function $\cos_{M,\epsilon} \in \sum^1(\Psi)$ such that

$$\sup_{\lambda \in [-M, M]} |\cos_{M,\epsilon}(\lambda) - cos(\lambda)| < \epsilon.$$

Proof: Let F be the cosine squasher previously defined. By adding, subtracting, and scaling a finite number of affinely shifted versions of F, we can get the cosine function on any interval $[-M, M]$. Since F is continuous, we may apply Lemma 2 and the triangle inequality to easily obtain the result. Indeed, let G be an element of $\sum^1(\Psi)$. We then have on the interval $[-M, M]$,

$$\begin{aligned} |G(\lambda) - cos(\lambda)| &\leq |G(\lambda) - F(\lambda)| + |F(\lambda) - cos(\lambda)| \\ &= |G(\lambda) - F(\lambda)| + 0 \\ &< \epsilon \end{aligned}$$

where the last line followed from Lemma 2.

Lemma 4: Let $g(\cdot) = \sum_{j=1}^{Q} \beta_j \cos(A_j(\cdot)), A_j \in A$. For arbitrary squashing function Ψ, arbitrary compact $C \subset I\!R$, and for arbitrary $\epsilon > 0$, there is an $f \in \sum^1(\Psi)$ such that $\sup_{x \in C} |g(x) - f(x)| < \epsilon$.

Proof: Pick $M > 0$ such that for $j \in \{1, 2, \ldots, Q\}$, $A_j(C) \subset [-M, M]$. Because Q is finite, C is compact and the $A(\cdot)$ are continuous, such an M can be found. Let $Q_1 = Q \cdot \sum_{j=1}^{Q} |\beta_j|$. From Lemma 3, for all $x \in C$ we have $|\sum_{j=1}^{Q} \beta_j \cos_{M,\epsilon}(A_j(x)) - g(x)| < \epsilon$. Because $\cos_{M,\epsilon/Q} \in \sum^1(\Psi)$, we see that $f(\cdot) = \sum_{j=1}^{Q} \cos_{M,\epsilon/Q}(A_j(\cdot)) \in \sum_1(\Psi)$.

Now we turn to proving the theorem. By Lemma 1, the trigonometric polynomials $\{\sum_{j=1}^{Q} \beta_j \prod_{k=i}^{l_j} \cos(A_{jk}(\cdot)) : Q, l_j \in I\!N, \beta_j \in I\!R, A_{jk} \in A\}$ are uniformly

dense on compact sets in $C(\mathbb{R})$. By repeated application of the trigonometric identity $\cos(a)\cos(b) = \cos(a+b) - \cos(a-b)$, we may write every trigonometric polynomial in the form $\sum_{t=1}^{T} \alpha_t \cos(A_t(\cdot))$ where $\alpha_t \in \mathbb{R}$ and $A_t \in A$. The desired result now follows from Lemma 4.

This now gives us another class of acceptable functions for the u_j in Figure 6, and choosing a squashing function will ensure that the output of each u_j is always between 0 and 1.

Radial Basis Function Networks

An important type of function that may be used in some classifying networks is the radial basis function, and more specifically, the Gaussian basis function. While we cannot generalize that in all cases a basis function network may be used for uniform approximation, there are some examples that are useful. Information about the universal approximation capability of radial basis function networks may be found in [17]. We define a radial basis function as a function which depends only on the norm of the argument. In other words, if f is a radial basis function and $\|x\| = \|y\|$, then $f(x) = f(y)$.

We now give an example of a case when uniform approximation is possible using a radial basis function network. In this particular instance the basis functions are Gaussian, functions that have other useful properties for approximation networks. Let H be a Hilbert space with inner product $< \cdot, \cdot >$ and norm $\| \cdot \|$ defined in the usual way. We are interested mainly in $H = \mathbb{R}^n$ with $\|x\| = \sum_j x_j^2$. Let $C \subset H$ be compact and let $V \subset H$ be nonempty, con-

vex, and satisfy the condition that for x_1, $x_2 \in C$ with $x_1 \neq x_2$ there exists $u \in V$ such that $\|x_1 - u\| \neq \|x_2 - u\|$. We can, for example, take V to be C as long as C is convex, or we can take V to be any nonempty convex subset of H containing an interior point. Let P be a nonempty subset of $(0, \infty)$ or $(-\infty, 0)$ that is closed under addition. Finally, let $L = \{g : C \mapsto I\!R : g(x) = \sum_{j=1}^{m} a_j \exp(-\alpha_j \|x - v_j\|^2), m < \infty, a_j \in I\!R, \alpha_j \in P, v_j \in V$. It is immediately seen that the structure of L is of the form needed for the elements u_j in Figure 6. With these assumptions we get the following theorem.

Theorem 18 Let $f : C \mapsto I\!R$ be continuous and let $\epsilon > 0$. Then there exists a $g \in L$ such that

$$|f(\alpha) - g(\alpha)| < \epsilon, \alpha \in C.$$

Proof: Using the property above and the convexity of V, we see that given a_1, $a_2 \in I\!R$, α_1, $\alpha_2 \in P$, and v_1, $v_2 \in V$

$$a_1 \exp(-\alpha_1 \|x - v_1\|^2) a_2 \exp(-\alpha_2 \|x - v_2\|^2) = b \exp(-(\alpha_1 + \alpha_2) \|x - w\|^2)$$

for some $b \in I\!R$ and $w \in V$. Also we can see that $\alpha_1 + \alpha_2 \in P$. So L is an algebra. Choose x_1 and x_2 in C and assume that $x_1 \neq x_2$. Then $\|x_1 - v\| \neq \|x_2 - v\|$ for some $v \in V$ by our first assumption.. Therefore $\exp(-\alpha \|x_1 - v\|) \neq \exp(-\alpha \|x_2 - v\|)$ so L separates the points of C. Therefore, by the Stone-Weierstrass theorem, the proof is complete.

Thus, in this somewhat less general compact space, the Gaussian basis functions are capable of uniformly approximating any continuous function in

$\mathbb{R}$. They therefore may be used as the elements u_j in our original network.

7. Applications

Classifier Example

At this point we are ready to give an example of an actual classification network using the framework that we have provided. This example will also show how the mathematical formulations that we have been making relate to the problems related to signal classification that were initially discussed.

Let X be the space of continuous real-valued functions defined on $[0,1]^n$ with $\|\cdot\|$ the usual sup norm. Let k and r be positive constants and let $\mathrm{Lip}(k)$ denote the subset of X consisting of the elements of X that satisfy a Lipschitz condition: $|x(a) - x(b)| \leq k|a - b|$ for all a and b. This is a typical way to deal with a good class of nonlinear functions. Let $x, x_2, \ldots, x_m$ be distinct elements of $\mathrm{Lip}(k)$ and let $C_j = \{x \in \mathrm{Lip}(k)\colon \|x - x_j\| \leq r\}$ for each $j = 1, 2, \ldots, m$.

Now assume that $r < (1/2)\min_{i \neq j} \|_i - x_j\|$. It is clear that the C_j are pairwise disjoint if this condition is satisfied. Since each C_j is a closed bounded subset of X that is equicontinuous on $[0,1]^n$, we get a result thanks to the Arzela-Ascoli theorem (see [15]) showing that the C_j are compact. As we have shown earlier, since the C_j are compact and pairwise disjoint, the union $\bigcup_j C_j$ is also compact.

We now introduce a theorem in [20] without the proof given there.

Theorem 19 Let X denote the normed linear space of $\mathbb{R}$-valued continuous functions on $\mathcal{I} := [0,1]^n$, with the usual max norm. Let $g \in X^*$, and let $\epsilon > 0$. Then there are points $\alpha_1, \ldots, \alpha_p \in \mathcal{I}$, points $c_1, \ldots, c_p \in \mathbb{R}$, and a $q \in X$ such

that

$$\sup_{x \in C} |g(x) - \sum_{j=1}^{P} c_j x(\alpha_j)| < \epsilon$$

and

$$\sup_{x \in C} |g(x) - \int_{\mathcal{I}} q(\alpha) x(\alpha) d\alpha| < \epsilon.$$

This theorem shows that a classifier can be found in this case using a simple sampling and summing operation or an integration. It applies directly to our example at hand since we are working on $[0, 1]^n$. We now know that it is possible to classify the signals in our example using the structure in Figure 6 where the functional y_j performs the sampling and summing or integration operation

This problem is very applicable to the examples discussed earlier. If $n =$ 1, 2, or 3, we are classifying continuous signals in one, two, or three variables. This is the kind of sensor input that we might have in the automatic target identification and pattern recognition examples that were mentioned earlier.

Conclusions

We have described a specific neural network structure that is capable of solving certain classification problems. This structure has the form of a single hidden layer feedforward neural network and therefore possesses the advantages of neural networks that were mentioned above. It has a simple framework that is easily built in hardware or simulated in software.

It is important to note that there are limitations to the methods presented here. All of the proofs are existence proofs. They guarantee that a

solution is possible and in some cases give a general idea on how it might be accomplished. For example, we have seen how certain classes of functions such as sigmoids and polynomials are capable of being used as the activation functions (the u_j) in a classifying neural network. What has not been determined is the number of nodes needed. We can only say that classification is possible with a finite number of nodes. Further, we have not given a certain method of finding the weights c_j in Figure 6. This is typically what we referred to as training the neural network.

In spite of these shortcomings, we have succeeded in providing a general framework capable of studying the important problem of signal classification. We have accomplished this by using well-known theorems dealing with approximation. This area of research is fairly new and has proven extremely useful so far, and interest in it will continue to grow in the future.

Bibliography

[1] J.-P. Antoine and P. Vandergheynst, "Target detection and recognition using two-dimensional isotropic and anisotropic wavelets," *Fifth International Conference on Automatic Object Recognition*, pp. 20-31, 1995.

[2] G. Bachman and L. Narici, *Functional analysis*, New York, Academic Press, 1966.

[3] D. Burr, "Artificial neural networks: a decade of progress," *Artificial Neural Networks for Speech and Vision*, Chapman and Hall, New York, 1994.

[4] E.W. Cheney, "Chapter 1: Ridge Functions," course notes for class *Approximation Theory*, Fall 1997.

[5] E.W. Cheney, "Chapter 2: Sigmoidal Ridge Functions," course notes for class *Approximation Theory*, Fall 1997.

[6] E.W. Cheney, "Chapter 25: Artificial Neural Networks," course notes for class *Approximation Theory*, Fall 1997.

[7] Charles K. Chui and Xin Li, "Approximation by Ridge Functions and Neural Networks with One Hidden Layer," *Journal of Approximation Theory*, Vol. 70, pp. 131-141, 1992.

[8] G. Cybenko, "Approximation by Superpositions of a Sigmoidal Function," *Mathematics of Control, Signals, and Systems*, Vol. 2, No. 4, pp. 303-314,

1989.

[9] N. Dunford and J. Schwartz, *Linear Operators*, John Wiley and Sons, Inc., New York, 1988.

[10] A. R. Gallant and H. White, "There exists a neural network that does not make avoidable mistakes," *IEEE Second International Conference on Neural Networks*, pp. I:657-664, 1988.

[11] S. Haykin, *Neural Networks*, Macmillan College Publishing Company, New York, 1994.

[12] K. Hornik, M. Stinchcombe, and H White, "Multilayer Feedforward Networks are Universal Approximators," *Neural Networks*, Vol. 2, pp. 359-366, 1989.

[13] A. N. Kolmogorov and V. S. Fonin, *Introductory Real Analysis*, Dover Publications, Inc., New York, 1970.

[14] L. A. Lusternik and V. J. Sobolev, *Elements of Functional Analysis*, Frederick Ungar Publishing Co., New York, 1961.

[15] A. Mukherjea and K. Pothoven, *Real and Functional Analysis*, Frederick Ungar Publishing Co., New York, 1960.

[16] H. Niemann, *Pattern Analysis*, Springer-Berlag, New York, 1981.

[17] J. Park and I. W. Sandberg, "Universal Approximation Using Radial-Basis-

Function Networks," *Neural Computation*, Vol. 3, pp. 246-257, 1991.

[18] W. Rudin, *Functional Analysis*, McGraw Hill, New York, 1973.

[19] W. Rudin, *Principles of Mathematical Analysis*, Third Edition, McGraw Hill, New York, 1976.

[20] I. W. Sandberg, "General Structures for Classification," *IEEE Transactions on Circuits and Systems*, Vol. 41, No. 5, pp. 372-376, May 1994.

[21] G. M. Shepherd and C. Koch, "Introduction to synaptic circuits," *The Synaptic Organizations of the Brain*, Oxford University Press, New York, 1990.

[22] B. W. Stiles, I. W. Sandberg, and J. Ghosh, "Complete Memory Structures for Approximating Nonlinear Discrete-Time Mappings," *IEEE Transactions on Neural Networks*, Vol. 8, No. 6, pp. 1397-1409, 1997.

[23] M. H. Stone, "A Generalized Weierstrass Approximation Theorem," *Studies in Mathematics*, Prentice Hall Inc., Englewood Cliifs, NJ, 1962.

[24] X. Sun and E. W. Cheney, "The fundamentality of sets of ridge functions," *Aequationes Mathematicae*, Vol. 44, pp. 226-235, 1992.

VITA

William Christopher Pritchett was born in Houston, Texas on September 1, 1973, the son of Carol Coulter Pritchett and Joseph Lucien Pritchett, III. After completing his work at Highland Park High School, Dallas, Texas, in 1992, he entered The United States Naval Academy in Annapolis, Maryland. He received the degree of Bachelor of Science from The United States Naval Academy in May, 1996. Additionally, he was commissioned as an Ensign in the United States Navy in May, 1996. In August, 1996, he entered The Graduate School at The University of Texas.

Permanent Address: 1825 Wroxton Rd
Houston, Texas 77005

This thesis was typed by the author.

20 SINPS Tn 1100
1/99 22527-157

DUDLEY KNOX LIBRARY
3 2768 00360688 0