PERGAMON

Neural Networks 12 (1999) 467–477

Neural Networks

Contributed article

Neural network approximation of continuous functionals and continuous functions on compactifications

M.B. Stinchcombe*

Department of Economics, The University of Texas at Austin, Austin, TX 78712-1173, USA

Received 21 July 1997; received in revised form 23 July 1998; accepted 23 July 1998

Abstract

This article characterizes the set of activation functions, bounded or unbounded, that allow feedforward network approximation of the continuous functions on the classic two-point compactification of $\mathbb{R}^1$. The characterization fails when the set of targets are continuous functions on the classic compactifications of $\mathbb{R}^n$, $n \geq 2$. Nonpolynomial, analytic activation functions with input-to-hidden weights in very limited sets allow approximation of continuous function over compact sets in $\mathbb{R}^n$, while even sigmoidal activation functions with weights in limited sets cannot approximate continuous functions on compactifications. The abstract structure foregrounded by compactification leads directly to possibility results for multi-layer networks and possibility results for neural networks in infinite dimensional settings.

Keywords: Activation functions; Compactification; Feedforward network approximation

1. Introduction

There are many different proofs of the ability of single hidden layer feedforward (SLFF) networks to approximate any continuous functions over arbitrary compact subsets of $\mathbb{R}^n$ for arbitrary, finite n.[1] Loosely, these results are called 'possibility theorems' – they show what is possible with SLFF networks. *Im*possibility results are useful too – by showing what is not possible with SLFF networks, they help clarify what kinds of applications will not work.

The range of impossibility results is immense. One starting point is the observation that it is impossible to continuously parameterize neural networks and approximate all continuous bounded functions uniformly. This is because the set of continuous bounded functions is not sup norm separable while the parameter sets for neural networks are. As an example of this logic, no SLFF network based on (say) a sigmoid can approximate (say) the sine function over all of $\mathbb{R}^1$. If capturing sinusoidal limit behavior is important, typical SLFF networks are not the right tool.[2]

While coming arbitrarily close to all continuous bounded functions over $\mathbb{R}^1$ or $\mathbb{R}^n$ is impossible, perhaps it is still possible to come arbitrarily close to a smaller set of targets. This article begins by considering as targets the continuous functions on $\mathbb{R}^1$ and $\mathbb{R}^n$ that have good limit behavior at infinity. Good limit behavior at infinity is equivalent to being a continuous function on an appropriate compactification of $\mathbb{R}^1$ or $\mathbb{R}^n$. Parts of this topic were studied in Chen et al. (1995) and Huang and Babri (1997). The work here shows that, to some extent, the possibility results carry over to this setting; the impossibility results provide the qualifications.

The second theme in this article arises from seeing the targets as continuous functions on compactifications. The abstract structure foregrounded by the compactifications leads to extensions of the possibility results to infinite dimensional contexts.

The next section gives the definitions and results for the 1-dimensional case and the following gives the same for the

* Tel.: + 1-512-471-3211; fax: + 1-512-471-3510.

E-mail address: maxwell@mundo.eco.utexas.edu (M.B. Stinchcombe)

[1] See e.g. Gallant and White (1988), Irei and Miyake (1988), Funahashi (1989), Cybenko (1989), Hecht-Nielson (1989), Hornik et al. (1989), Chen et al. (1990), Chui and Li (1992). Hornik et al. (1990), Cardaliaguet and Euvrard (1992), Li (1996) show that smooth SLFF networks can simultaneously approximate a function and its derivatives. Hornik et al. (1990), Stinchcombe and White (1989, 1990, 1992), Hornik (1991), Hornik (1993), and Leshno et al. (1993) extended these results in various directions, using different metrics, loosening the restrictions on the activation function and using limited sets of input-to-hidden weights. Hornik et al. (1989) showed that feedforward networks with $k \geq 1$ sigmoidal layers have this property, Ito (1994) examined this and deeper properties of feedforward networks with $k = 2$ sigmoidal layers.

[2] Chui et al. (1996) provide impossibility results of an entirely different form-sharp lower bounds on the rates of approximation achievable by SLFF networks.

0893-6080/99/$ - see front matter
PII: S0893-6080(98)00108-7

n-dimensional case, $n \geq 2$, n finite. Theorems 2.1 and 2.1′ characterize those activation functions, bounded or not, that allow feedforward approximation of the continuous functions on the classic two-point compactification of $\mathbb{R}^1$. Theorem 2.2 gives more detail about neural networks based on activation functions that fail the characterization. If uniform approximation over compact sets rather than compactifications is desired, then Theorem 2.3 shows that it is possible to use very limited sets of input-to-hidden weights. Theorem 2.4 shows that the limited sets of weights results do not carry over to approximating functions on the classic two-point compactification. Theorems 3.1, 3.1′ and 3.2 show that the characterization results for one dimension do not carry over to higher dimensions. Theorems 3.3 and 3.4 give parallel possibility and impossibility results for SLFF networks with limited weights. Section 4 briefly discusses how these results relate to radial basis function networks. The following section, Section 5, takes up the second theme, discussing some lessons about the structure of input-to-hidden weight connections that come from the compactification perspective on limit behavior. Theorem 5.1 and its seven Corollaries cover multi-layer networks, neural network approximation results in infinite dimensional settings, as well as a separate and more potentially more fruitful approach to approximating functions on compactifications. The penultimate section, Section 6, discusses how this work relates to the previous work on this topic, while Section 7 contains the proofs.

2. Approximation by SLFF networks in one dimension

This section begins with the needed definitions and notations, and then turns to the results in one dimension.

2.1. Definitions and notation

Definitions for spaces of continuous functions precede those for spaces of measurable functions; those for SLFF networks and denseness follow.

2.1.1. Continuous functions

The space C^1 is the set of continuous functions on $\mathbb{R}^1$. For present purposes, two different metrics will be used. First, the extended real-valued sup norm metric,

$$d_\infty(f,g) = \sup_{x\in\mathbb{R}} |f(x) - g(x)|. \tag{1}$$

Note that $d_\infty(f,g) = +\infty$ is possible if either f or g is unbounded. However, with the convention that $r + \infty = +\infty$ for any $r \in [0,\infty]$, $d_\infty(\cdot,\cdot)$ satisfies the three defining properties of a metric. The second metric gives uniform convergence over compact subsets,

$$d_C(f,g) := \sum_{n\in\mathbb{N}} 2^{-n} \min\{\sup_{|x|\leq n} |f(x) - g(x)|, 1\}. \tag{2}$$

With the metric d_C, the space C^1 and all of its vector subspaces are topological vector spaces, while the vector subspaces of $C_b^1 \subset C^1$, the set of bounded elements of C^1, are topological vector spaces with the metric $d\infty$. Note that d_∞ is the tighter metric; $d_\infty(f^\alpha, f) \to 0$ implies that $d_C(f^\alpha, f) \to 0$, so that d_∞-denseness is a stronger property.

Two special vector subspaces of C^1 will play a large role in what follows. The space $\vec{C}^1 \subset C^1$ is the set of continuous functions having real limits at $\pm\infty$. Formally,

$$\vec{C}^1 = \left\{ f \in C^1 : \lim_{x\to-\infty} f(x) \text{ and } \lim_{x\to+\infty} f(x) \text{ exist in } \mathbb{R}^1 \right\}. \tag{3}$$

Note that $\vec{C}^1$ is a vector subspace of C_b^1. The space $\mathring{C}^1 \subset \vec{C}^1$ is the set of continuous functions having the same limit at $\pm\infty$,

$$\mathring{C}^1 = \left\{ f \in \vec{C}^1 : \lim_{x\to-\infty} f(x) = \lim_{x\to+\infty} f(x) \right\}. \tag{4}$$

From an abstract point of view, the space $\mathring{C}^1$ is the set of continuous functions on $\mathbb{R}^1$ having a unique continuous extension to the classical one-point compactification of $\mathbb{R}^1$ (e.g. by stereographic projection as in Guillemin and Pollack (1974), p. 12)). In the same way, $\vec{C}^1$ is the set of continuous functions on $\mathbb{R}^1$ having a unique continuous extension to the two-point compactification of $\mathbb{R}^1$, that is, to $[-\infty, +\infty]^1$ (e.g. if Φ is the c.d.f. of a smooth random variable with everywhere positive density, then Φ gives a homeomorphism of $(-\infty,+\infty)$ with (0,1), defining $\Phi(-\infty) = 1 - \Phi(+\infty) = 0$ completes the definition). For general treatments of compactifications, see e.g.Kelley (1955), Ch. 5) or Hurd and Loeb (1985), Section III.7). As the one- and two-point compactifications are homeomorphic to a circle and a compact interval respectively, the sets $\mathring{C}^1$ and $\vec{C}^1$ are isometrically isomorphic to sets of continuous functions on these compact sets.

2.1.2. Measurable functions

The measurable versions of these spaces replace ''C'' with ''M''. Specifically, the space M^1 is the set of (Borel) measurable functions on $\mathbb{R}^1$ with the metrics d_∞ and d_C defined as before. The space $M_b^1 \subset M^1$ is the set of bounded elements of M^1. M_b^1 and its vector subspaces, e.g. C_b^1, are topological vector spaces with the metric d_∞; vector subspaces of M^1 containing only locally bounded functions are topological vector spaces with the metric d_C.

The space $\vec{M}^1 \subset M^1$ is the set of measurable functions having limits at $\pm\infty$. Formally, this is

$$\vec{M}^1 = \left\{ f \in M^1 : \lim_{x\to-\infty} f(x) \text{ and } \lim_{x\to+\infty} f(x) \text{ exist in } \mathbb{R}^1 \right\}. \tag{5}$$

The set $\vec{M}_b^1$ is the set of bounded elements in $\vec{M}^1$. The sets $\mathring{M}^1$ and $\mathring{M}_b^1$ are the subsets of $\vec{M}^1$ and $\vec{M}_b^1$ (respectively) that have the same limit at $\pm\infty$.

2.1.3. SLFF networks and denseness

At issue is the ability of SLFF networks to approximate elements of subsets of C^1 in various metrics, especially d_∞ and d_C.

Definition 2.1. For $g \in M^1, S_g^1$ denotes the set of **SLFF networks on $\mathbb{R}^1$ based on (the activation function)** g, that is, the set of functions of the form $h(x) = \sum_{j=1}^{J} \beta_j g(w_j x + b_j)$ where $J \in \mathbb{N}$, $\beta_j \in \mathbb{R}^1$ *and* $(w_j, b_j) \in \mathbb{R}^2$.

Definition 2.2. For $S, C \subset M^1$ and for a metric d on M^1, S *is* d-**outside dense for** C if the d-closure of S contains C, and S *is* d-**inside dense for** C if the d-closure of $S \cap C$ contains C.

Note that if S is d-inside dense for C, then it must be d-outside dense for C.

Of particular interest are generalized sigmoid activation functions.

Definition 2.3. A function in $\vec{M}_b^1 \backslash \mathring{M}_b^1$ is called a **generalized sigmoid**; a function in $\vec{C}^1 \backslash \mathring{C}^1$ is called a **continuous generalized sigmoid**.

In terms of the definitions used here, Cybenko's result (Cybenko, 1989) is that S_g^1 is d_C-outside dense for C^1 if g is a continuous generalized sigmoid (though nothing in the proof requires changing if g is a generalized sigmoid). Funahashi (1989) and Hornik et al. (1989) came to the same conclusion for non-decreasing generalized sigmoids. Chen et al. (1990) give a constructive proof of Cybenko's (Cybenko, 1989) result. If g is a continuous generalized sigmoid, then S_g^1 is d_C-inside dense for C^1. As the example in the previous section noted, if g is a continuous generalized sigmoid, S_g^1 can be continuously parameterized by a separable set, implying that S_g^1 cannot be d_∞-outside dense for any non-separable subset of M^1 or C^1.

For $g \in M^1$ and $(w, b) \in \mathbb{R}^2$, define $g_{(w,b)} \in M^1$ by $g_{(w,b)}(x) = g(wx + b)$. For $\mathcal{O} \subset \mathbb{R}^2$, let $F_g^1(\mathcal{O}) = \{g_{(w,b)} : (w, b) \in \mathcal{O}\}$, and let $S_g^1(\mathcal{O}) = \mathit{span}\, F_g^1(\mathcal{O})$. General properties of the sets $F_g^1(\mathcal{O})$ and their implications for neural networks can be found in Stinchcombe (1995). With this notation, $S_g^1(\mathbb{R}^2) = \mathit{span}(F_g(\mathbb{R}^2)) = S_g^1$.

Definition 2.4. For $C \subset M^1$ and for a metric d on M^1, SLFF networks based on g are ***d*-outside (respectively *d*-inside)** *dense for* C **with limited weights** if the d-closure of $S_g^1(\mathcal{O})$ (**respectively** $S_g^1(\mathcal{O}) \cap C$) contains C for every non-empty open $\mathcal{O} \subset \mathbb{R}^2$.

Many of the proofs of inside or outside denseness of SLFF networks require that the input-to-hidden weights, w and b as mentioned before, be allowed to be arbitrarily large. This is difficult to implement in either hardware or software, and not credible for natural networks. Networks that are dense with limited weights are correspondingly easier to implement and more credible as models of natural processes. Stinchcombe and White (1990) give conditions on g for which $S_g^1(\mathcal{O})$ is d_C-inside dense for C^1 when $\mathcal{O}$ is a bounded set containing the origin. Stinchcombe and White (1992) give conditions on g (satisfied by e.g. the p.d.f. or the c.d.f. of a logistic or Gaussian random variable) that are sufficient to imply that SLFF networks based on g are d_C-inside dense for C^1 with limited weights.

2.2. Results in one dimension

The characterization results precede the discussion of limited weights.

2.2.1. Characterization

The inside version of the characterization result is

Theorem 2.1. *S_g^1 is d_∞-inside dense for $\vec{C}^1$ if and only if S_g^1 contains a continuous, generalized sigmoid.*

The outside version of this result is

Theorem 2.1′. *S_g^1 is d_∞-outside dense for $\vec{C}^1$ if and only if the d_∞-closure of S_g^1 contains a continuous, generalized sigmoid.*

Note that if S is d_∞-outside (inside) dense for $\vec{C}^1$, then S is d_C-outside (inside) dense for C^1 because d_∞ is the tighter metric and continuous functions on closed sets have continuous extensions.

Many unbounded continuous g's have the property that S_g^1 contains a continuous generalized sigmoid. For example, Hornik et al. (1990) set $g(x) = \int_{-\infty}^{x} h(t)\,\mathrm{d}t$ where h is the cumulative distribution function of any random variable with finite expectation. The function g is unbounded, yet $\eta(x) = g(x + 1) - g(x)$ is continuous, $\lim_{x \to -\infty} \eta(x) = 0$, and $\lim_{x \to +\infty} \eta(x) = 1$. Mhaskar and Micchelli (1992), and Mhaskar (1993) show that SLFF networks based on (necessarily unbounded) k'th order sigmoidal activation functions[3] are d_C-outside dense for C^η.

There are bounded and unbounded continuous functions for which the intersection of $\vec{C}^1$ and the d_∞-closure of S_g^1 contains only the constant functions. For unbounded functions, let $g(x)$ be any non-constant polynomial; for a bounded non-constant function, let $g(x) = \sin(xe^{x^2})$.

The first impossibility result gives more detail about the failure of S_g^1 to be dense.

Theorem 2.2. *If $g \in \mathring{M}_b^1$ (respectively $\mathring{C}^1$), then the d_∞-closure of S_g^1 is a subset of $\mathring{M}_b^1$ (respectively $\mathring{C}^1$).*

[3] A function g is a kth order sigmoidal function if (a) it is bounded by a polynomial of degree at most k on $\mathbb{R}$, and (b) the limit of $x^{-k}\sigma(x)$ is 0 as x goes to $-\infty$, and is 1 as x goes to ∞.

This implies that SLFF networks based activation functions in $\mathring{M}_b^1$ cannot uniformly approximate anything outside of $\mathring{M}_b^1$, e.g. nothing in $\vec{C}^1 \backslash \mathring{C}^1$.

2.2.2. Limited weights

The characterization/possibility result for limited weights is

Theorem 2.3. *If $g \in C^1$ is analytic, then SLFF networks based on g are d_C-inside dense for C^1 with limited weights if and only if g is not a polynomial.*

If g is the density or c.d.f. of a Gaussian or a logistic random variable, then Stinchcombe and White (1992) showed that SLFF networks based on g are d_C-inside dense for C^1 with limited weights. For a smooth function g, let A_g denote the set of points at which g fails to be analytic. I conjecture that if g is smooth, SLFF networks based on g are d_C-inside dense for C^1 with limited weights if and only if the closure of A_g has empty interior and g is not equal to a polynomial on any open set.

The impossibility result for limited weights is

Theorem 2.4. *If $g \in \vec{M}^1$, then SLFF networks based on g are never d_∞-outside dense for $\vec{C}^1$ with limited weights.*

To the extent that networks with limited weights are more credible, this result means that SLFF networks based on generalized sigmoids are not the right tool for uniformly capturing the restricted limit behavior of functions in $\vec{C}^1$.

3. Approximation by SLFF networks in many dimensions

Definitions and notation parallel the previous section and precede the results.

3.1. Definitions and notation

Definitions for spaces of continuous functions precede those for spaces of measurable functions; those for SLFF networks and denseness follow.

3.1.1. Continuous functions

The space C^n is the set of continuous functions on $\mathbb{R}^n$. For present purposes, two different metrics will be used. The first is the extended real-valued sup norm metric, $d_\infty(f,g) = \sup_{\mathbf{x}\in\mathbb{R}^n}|f(\mathbf{x}) - g(\mathbf{x})|$. The second gives uniform convergence over compact subsets, $d_C(f,g) := \sum_{n\in\mathbb{N}} 2^{-n}\min\left\{\sup_{|\mathbf{x}|\le n}|f(\mathbf{x}) - g(\mathbf{x})|, 1\right\}$. With the metric d_C, the space C^n and all of its vector subspaces are topological vector spaces, while the vector subspaces of $C_b^n \subset C^n$, the set of bounded elements of C^n, are topological vector spaces with the metric d_∞.

Two special vector subspaces of C_b^n will play a large role in what follows. $\vec{C}^n$ is the set of continuous functions with real limits at all points in $[-\infty, +\infty]^n$.

Definition 3.1. A sequence $x^\alpha = (x_1^\alpha, \ldots, x_n^\alpha)$ in $\mathbb{R}^n$ *converges in* $[-\infty, +\infty]^n$ if for each coordinate $i \in \{1, \ldots, n\}$, there is an extended real number $x_i^o \in [-\infty, +\infty]$ such that $x_i^\alpha \to x_i^o$.

Formally,

$$\vec{C}^n = \{f \in C^n : \text{ if } x^\alpha \text{ converges in } [-\infty,+\infty]^n, \text{ then } \lim_\alpha f(x^\alpha) \text{ exists in } \mathbb{R}^1\}. \tag{6}$$

The space $\mathring{C}^n \subset \vec{C}^n$ is the set of continuous functions having the same limit at all infinite points in $[-\infty,+\infty]^n$.

$$\mathring{C}^n = \{f \in \vec{C}^n : \text{ if } \|x^\alpha\|, \|y^\alpha\| \to \infty, \text{ then } \lim_\alpha f(x^\alpha) = \lim_\alpha f(y^\alpha) \text{ in } \mathbb{R}^1\} \tag{7}$$

The space $\mathring{C}^n$ can be identified with the set of continuous functions on $\mathbb{R}^n$ having a unique continuous extension to the classical 1-point compactification of $\mathbb{R}^n$, and $\vec{C}^n$ is the set of continuous functions on $\mathbb{R}^n$ having a unique continuous extension to the $2n$-point compactification of $\mathbb{R}^n$, that is, to $[-\infty, +\infty]^n$. As the 1- and $2n$-point compactifications are homeomorphic to a sphere and a compact product of intervals respectively, the sets $\mathring{C}^1$ and $\vec{C}^1$ are isometrically isomorphic to sets of continuous functions on these compact sets.

3.1.2. Measurable functions

The measurable versions of these spaces are $M^n, M_b^n, \vec{M}^n, \vec{M}_b^n, \mathring{M}^n,$ *and* $\mathring{M}_b^n$, with definitions that parallel those in Section 2.1.2.

3.1.3. SLFF networks and denseness

At issue is the ability of SLFF networks to approximate elements of subsets of C^n in various metrics, especially d_∞ and d_C. The definitions of d-inside and d-outside denseness are exactly as mentioned earlier.

Definition 3.2. For $g \in M^1, S_g^n$ **denotes the set of** *SLFF* **networks on $\mathbb{R}^n$ based on (the activation function)** g, that is, the set of functions of the form $h(\mathbf{x}) = \sum_{j=1}^J \beta_j g(w_j \cdot \mathbf{x} + b_j)$ where $J \in \mathbb{N}$, $\beta_j \in \mathbb{R}^1, (w_j, b_j) \in \mathbb{R}^{n+1}$, and $w_j \cdot \mathbf{x}$ is the dot product of the n-vectors w_j and $\mathbf{x}$.

For $g \in M^1$ and $(w,b) \in \mathbb{R}^{n+1}$, define $g_{(w,b)} \in M^1$ by $g_{(w,b)}(\mathbf{x}) = g(w \cdot \mathbf{x} + b)$. For $\mathcal{O} \subset \mathbb{R}^{n+1}$, let $F_g^n(\mathcal{O}) = \{g_{(w,b)} : (w,b) \in \mathcal{O}\}$, and let $S_g^n(\mathcal{O}) = \mathit{span}\, F_g^n(\mathcal{O})$.

Definition 3.3. For $C \subset M^n$ and for a metric d on M^n, SLFF networks based on g are d-**outside (respectively** d-**inside) dense for** C **with limited weights** if the d-closure of

$S_g^n(\mathcal{O})$ (respectively $S_g^n(\mathcal{O}) \cap C$) contains C for every non-empty open $\mathcal{O} \subset \mathbb{R}^{n+1}$.

The discussion of the desirability of good approximation properties with limited weights given earlier carries over directly to the case of many dimensions.

3.2. Results in two or more dimensions

For $n \geq 2$, inside and the outside approximation are very different. The discussion of this difference precedes the discussion of approximation with limited weights.

3.2.1. Inside vs. outside

Recall that Theorem 2.1 asserts that S_g^1 is d_∞-inside dense for $\vec{C}^1$ if and only if S_g^1 contains a continuous, generalized sigmoid. The next result proves that inside approximation of $\mathring{C}^n$, $n \geq 2$, is not feasible with SLFF networks based on sigmoids – the most that can be approximated are the constant functions. Theorem 3.2 shows that inside approximation of $\vec{C}^n$, $n \geq 2$, is not feasible for SLFF networks – the most that can be approximated are functions with zero cross effects.

Theorem 3.1. *For all $n \geq 2$, if $g \in \vec{C}^1 \backslash \mathring{C}^1$, then $S_g^n \cap \mathring{C}^n$ is a subset of the constant functions.*

Compare the following with Theorems 2.1′ and 2.2.

Theorem 3.1′. *For all $n \geq 2$, the d_∞-closure of S_g^1 containing a generalized sigmoid is sufficient but not necessary for S_g^n to be d_∞-outside dense for $\mathring{C}^n$.*

The sharp distinction between inside and outside approximation for $n \geq 2$ is somewhat troublesome. The classes of functions being approximated are those with good limit behavior at infinity. Inside approximation asks that the approximants be members of the class being approximated. If this limit behavior at infinity is important, then it may seem peculiar to use approximants that do not have it as in Theorem 3.1′. Another cost is revealed in the methods of proof. From the proof of Theorem 3.1 we have

Lemma B. *If $g \in \vec{C}^1$ is not constant and $f \in S_g^n \cap \vec{C}^n$, then f can be expressed as a finite linear combination $f = \sum_{j=1}^J \beta_j g_{(w_j,b_j)}$ where each $w_j \in \mathbb{R}^n$ has at most one non-zero coordinate.*

The intuition behind Lemma B is easy to explain when $J = 1$ and $n = 2$. Suppose that $w = (1,1)$, and consider the function $f(x_1, x_2) = g_{(w,b)}(x_1, x_2) = g(x_1 + x_2 + b)$.. For this function to belong to $\vec{C}^2$, it must be the case that for all numbers r, $\lim_{\alpha \to \infty}(\alpha, r - \alpha)$ is the same because $\lim_{\alpha \to \infty} f(\alpha, r - \alpha) = (+\infty, -\infty)$ for all r. However, $f(\alpha, r - \alpha) \equiv g(r + b)$, so this requires that g does not depend on $r + b$, that is, this requires that g is constant.

The way to approximate an element $f \in \mathring{C}^n$ from the outside begins by giving an integral transform representation of a function φ that is d_∞-close to f, $\varphi(x) = \int g_{(w,b)}(\mathbf{x})\, d\nu(\mathbf{x})$ where ν is a finite signed measure. One then appeals to convergence results for integrals to show that for large J, there are sums of the form $\sum_{j=1}^J \beta_j g_{(w_j,b_j)}(\mathbf{x}) \simeq \int g_{(w,b)}(\mathbf{x})\, d\nu(\mathbf{x})$. In other words, one approximates ν by the measure that puts mass β_j on the point (w_j, b_j), $j = 1, \ldots, J$. What Theorem 3.1 requires of this representation is that the measure ν have no mass points (e.g. ν has a density with respect to Lebesgue measure on a sphere in Radon transformation proofs). In other words, for large J, no small subset of the β_j can be particularly large. In network terms, Lemma B requires that no small subset of the units in the network do very much of the work; that is, there can be no significant hidden-to-output parts of the network.

Combining Lemma B and Theorem 2.1 gives a characterization result for inside approximation by continuous networks in $\mathbb{R}^n$. The characterization involves functions on $\mathbb{R}^n$ that depend on only one coordinate. For $x = (x_1, \ldots, x_n) \in \mathbb{R}^n, i \in \{1, \ldots, n\}$, and $r \in \mathbb{R}^1$, define $(x,i,r) \in \mathbb{R}^n$ to be the vector with the number r replacing the i'th component of the vector x,

$$(\mathbf{x}, i, r) = (x_1, \ldots, x_{i-1}, r, x_{i+1}, \ldots, x_n) \in \mathbb{R}^n. \tag{8}$$

For $i \in \{1, \ldots, n\}$, the set of functions in $\vec{C}^n$ depending on only the i'th coordinate of their argument is

$$\mathcal{C}_i^n = \{f \in \vec{C}^n : \ (\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n)(\forall r \in \mathbb{R}^1)[f((\mathbf{x}, i, r)) = f((\mathbf{y}, i, r))]\}. \tag{9}$$

Each $\mathcal{C}_i^n$ is a closed linear subspace, as is their sum, $\mathcal{C}$,

$$\mathcal{C} = \sum_{i=1}^n \mathcal{C}_i^n = \left\{ f = \sum_{i=1}^n f_i : \ (\forall i \in \{1, \ldots, n\})[f_i \in \mathcal{C}_i^n] \right\}. \tag{10}$$

Note that a function in $\mathcal{C}$ can have no cross effects – if $f \in C$, then for all $\mathbf{x}, \mathbf{y}, r, s$,

$$f((\mathbf{x}, i, r)) - f((\mathbf{x}, i, s)) = f((\mathbf{y}, i, r)) - f((\mathbf{y}, i, s)). \tag{11}$$

If f is twice continuously differentiable, this is equivalent to $\partial^2 f / \partial x_i \partial x_j = 0, i \neq j$.

The next result is an immediate consequence of Lemma B. It shows that at best, only functions with no cross effects can be approximated from the inside using SLFF networks.[4]

Theorem 3.2. *For all n and for all $g \in C^1$, $S_g^n \cap \vec{C}^n \subset C$,*

[4] For example, consider the function $\varphi(x_1, x_2) = x_1 \cdot x_2 \cdot \psi(x_1, x_2)$ where ψ is a $C\infty$ function with values in [0,1] which is equal to 1 on the ball around 0 with radius 5, and which is equal to 0 outside the ball around 0 with radius 6. The function φ belongs to $\mathring{C}^2$ *limits*. An implication of Theorem 3.2 is that no SLFF network based on an activation function $g \in C^1$ has any element in $\vec{C}^2$ that is uniformly within 12.5 of φ.

and $S_g^n \cap \vec{C}^n$ is dense in $\mathscr{C}$ if and only if S_g^1 contains a continuous, generalized sigmoid.

In light of Theorem 3.2, Theorem 3.1 follows from the observation the intersection of $\mathring{C}^n$ and $\mathscr{C}$ contains only the constant functions.

For the later purposes of showing that feedforward networks with multiple hidden layers generally give inside approximations $\vec{C}^n$, Corollary 5.1.7, note that C, although limited, is a vector space of functions that separates points and contains the constant functions.

3.2.2. Limited weights

The characterization/possibility and impossibility results for limited weights are similar for general n.

Theorem 3.3. *If $g \in C^1$ is analytic, then SLFF networks based on g are d_C-inside dense for C^n with limited weights if and only if g is not a polynomial.*

Theorem 3.4. *If g is a generalized sigmoid and $n \geq 2$, then SLFF networks based on g are never d_∞-outside dense for $\mathring{C}^n$ with limited weights.*

As mentioned previously, to the extent that networks with limited weights are more credible, this result means that capturing the limit behavior of functions in $\mathring{C}^n$ using networks in $\vec{C}^n$ is not a reasonable hope for SLFF networks.

4. Radial basis function networks

Radial basis function[5] (RBF) networks on $\mathbb{R}^n$ are the span of functions of the form

$$G_{(A,c)}(x) = G((x-c)'A(x-c)) \tag{12}$$

where $x,c \in \mathbb{R}^n$, A is a $n \times n$ positive definite matrix, and $G: [0,\infty) \to \mathbb{R}$.

If G is continuous and $\lim_{r\to\infty} G(r)$ exists, as is usually assumed, then for all A and c, $G_{(A,c)} \in \mathring{C}^n$. This implies that the RBF networks can, at best, uniformly approximate functions in $\mathring{C}^n$, $n \geq 1$. For $n \geq 1$, Theorem 2.1 shows that SLFF networks can do better if e.g. g is a sigmoid. For $n \geq 2$, Theorem 3.1 shows that SLFF networks miss exactly the set that contains RBF networks when we ask that our approximants belong to the class being approximated. For $n \geq 2$, Theorem 3.1′ shows that SLFF networks can do much better than RBF networks provided that one allows for approximants outside the class being approximated.

[5] The denseness of RBF networks in d_C and other metrics has been investigated by Park and Sandberg, 1993, 1994, Elanayar and Shin (1994), Chen and Chen (1995a), Mhaskar and Micchelli (1995).

5. Inner functions, compositions, and compactification

The basic result in this section has implications for feedforward networks with multiple non-linear layers, for feedforward approximation of continuous functional on locally convex topological vector spaces, and for the successful inside approximation of functions with good limit behavior at infinity.

5.1. Basics

For any Hausdorff topological space X, $C(X)$ denotes the set of continuous functions on X. For any compact $K \subset X$ and (measurable) functions f, g on X, define $\rho_K(f,g) = \sup_{x\in X}|f(x)-g(x)|$. A set of measurable functions $\mathscr{A}$ on X is a **vector space** if for all $\alpha,\beta \in \mathbb{R}$ and all $f,g \in \mathscr{A}$, $\alpha f + \beta g \in \mathscr{A}$. The set $\mathscr{A}$ of functions on X **separates points** in X if for all $x \neq y$, $x,y \in X$, there exists an $a \in \mathscr{A}$ such that $a(x) \neq a(y)$. For a function $g: \mathbb{R} \to \mathbb{R}$ and a set of functions A on X, $g \circ A$ denotes the set of functions of the form $x \mapsto g(a(x)), a \in \mathscr{A}$, and $\mathtt{span}(g \circ \mathscr{A})$ denotes the $\mathtt{span}$ of $g \circ A$. If X is $\mathbb{R}^n$ and $\mathscr{A}$ is the set of functions $x \mapsto w\cdot x + b, w, x \in \mathbb{R}^n$, $b \in \mathbb{R}^1$, then $\mathtt{span}(g \circ \mathscr{A})$ is exactly S_g^n.

The basic result has an inside and an outside version.

Theorem 5.1. (a) *Suppose that S_g^1 is d_C-inside sense for C^1, that K is a compact subset of X, and that $\mathscr{A}$ is a vector space of measurable functions. If the ρ_K-closure of $C(K) \cap \mathscr{A}$ contains the constant functions and separates points in K, then $\mathtt{span}(g \circ \mathscr{A})$ is ρ_K-inside dense for $C(K)$.* (b) *Suppose that S_g^1 is d_C-outside for C^1, that K is a compact subset of X, and that $\mathscr{A}$ is a vector space of measurable functions. If the intersection of $C(K)$ and the ρ_K-closure of $\mathscr{A}$ contains the constant functions and separates points in K, then $\mathtt{span}(g \circ \mathscr{A})$ is ρ_K-outside dense for $C(K)$.*

5.2. Feedforward networks with many nonlinear layers

This Theorem extends approximation results for SLFF networks to approximation results for multiple layer feedforward networks. The set of feedforward networks on $\mathbb{R}^n$ with g being the activation function on 2 hidden layers is $S_{g,2}^n = span(g \circ S_g^n)$. Given that $S_{g,k-1}^n$ has been defined, $k \geq 3$, inductively define the set of feedforward networks with g being the activation function on k hidden layers by $S_{g,k}^n = span(g \circ S_{g,k-1}^n)$.

Corollary 5.1.1. *If S_g^n in d_C-inside dense (respectively d_C-outside dense) for $C(\mathbb{R}^n)$, then for all $k \geq 2$, $S_{g,k}^n$ is d_C-inside dense (respectively d_C-outside dense) for $C(\mathbb{R}^n)$.*

A more complicated proof of this can be found in Hornik et al. (1989).

5.3. Feedforward networks in infinite dimensional contexts

There has been some interest in the ability of feedforward

networks to approximate continuous non-linear functionals in infinite dimensional contexts.[6] The following direct corollaries to Theorem 5.1 generalizes the possibility results for feedforward networks in that literature.

Recall that all Banach spaces are locally convex topological vector spaces. The inside versions of the following three corollaries are also true.

Corollary 5.1.2. *Suppose that S_g^1 is d_C-outside dense for C^1 and that X is a locally convex topological vector space with dual space X^*. For all compact $K \subset X$, the class of functions of the form $\sum_{j=1}^{J} \beta_j g(l_j(x) + b_j)$, $J \in \mathcal{N}$, $l_j \in X^*$, $b_j \in \mathbb{R}^1$, is ρ_K-outside dense for C(K).*

A subset L of X^* is weak* dense in X^* if for all $x^* \in X^*$, all $x \in X$, and all $\varepsilon > 0$, there is an $l \in L$ such that $|x^*(x) - l(x)| < \varepsilon$.

Corollary 5.1.3. *The previous corollary remains true if the l_j are restricted to any weak*-dense subset of X^*.*

As an example, if X is (say) a separable L^2 space and $\{x_m : m \in N\}$ form a basis for X, then the l_j need only be of the form $\sum_{m=1}^{M} \beta_m x_m, M \in \mathcal{N}$.

Let $S_{g,1}^X$ denote the set of functions in either of the last two corollaries. Inductively define the multilayer feedforward networks with $k+1$ hidden layers by $S_{g,k+1}^X = \mathsf{span}(g \circ S_{g,k}^X)$.

Corollary 5.1.4. *Suppose that S_g^1 is d_C-outside dense for C^1 and that X is a locally convex topological vector with dual space X^*. For all compact $K \subset X$ and all integers k, the class of functions $S_{g,k}^X$ is ρ_K-outside dense for C(K).*

5.4. Functions with good limit behavior redux

The next results return to the approximation of functions with good limit behavior on $\mathbb{R}^n$. The basic observation is that if φ is a homeomorphism between compact sets K_1 and K_2, then the mapping $f \leftrightarrow f \circ \varphi$ from $C(K_2)$ to $C(K_1)$ is an isometric isomorphism between the sets of continuous functions on the two compact sets. In particular, this suggests that instead of looking for SLFF networks in $C \leftrightarrow^n$ (or $C\bigcirc^n$), one should look at networks of the form $\sum_{j=1}^{J} \beta_j g(w_j \cdot \varphi(\mathbf{x}) + b_j)$ where φ is a homeomorphism between the compactifications that form the domain of $\vec{C}^n$ (or $C\bigcirc^n$) and compact subsets of $\mathbb{R}^n$.

Recall that $\vec{C}^n$ is isometrically isomorphic to the set of continuous functions on $[-1,+1]^n$ and that $\mathring{C}^n$ is isometrically isomorphic to the set of continuous functions of S^n, the surface of the unit sphere in $\mathbb{R}^{n+1}$. Rather than insisting on using linear functions (unbounded and discontinous) on the domains of the functions $\vec{C}^n$ and $\mathring{C}^n$, the following two results translate the approximation problem back to more familiar territory.

Corollary 5.1.5. *Suppose that φ is a homeomorphism between $[-\infty,+\infty]^n$ and $[-1,+1]^n$, and that $\mathcal{A}$ is the set of affine functions on $[-1,+1]^n$. If S_g^n is d_C-inside dense (respectively d_C-outside dense) for C^n, then $\mathsf{span}(g \circ \mathcal{A} \circ \varphi)$ is d_∞-inside dense (respectively d_∞-outside dense) for $\vec{C}^n$.*

Corollary 5.1.6. *Suppose that φ is a homeomorphism between the one point compactification of $\mathbb{R}^n$ and S^{n+1} and that $\mathcal{A}$ is the set of affine functions on S^{n+1}. If S_g^n is d_C-inside dense (respectively d_C-outside dense) for C^n, then span $(g \circ \mathcal{A} \circ \varphi)$ is d_∞-inside dense (respectively d_∞-outside dense) for $\mathring{C}^n$.*

Provided that the affine structure of the functions inside the activation function is not crucial, these last two Corollaries suggest a method of approximation of any algebra of bounded functions: (1) If $\mathcal{F}$ is a supnorm closed algebra of bounded functions containing the constants, then it is isometrically isomorphic to the set of continuous functions on some compact Hausdorff space Y. (2) The space Y is homeomorphic to a compact subset K of $X = \mathbb{R}^{\mathcal{F}}$ with the product topology. The space X is a locally convex topological vector space. (3) Let φ be a homeomorphism between the domain of $\mathcal{F}$ and K. (4) The class of functions of the form $\sum_{j=1}^{J} \beta_j g(l_j(\varphi(x)) + b_j), l_j \in X^*$, will be ρ_K-dense. The steps (1)–(4) are the general form of the arguments used in the last two Corollaries.

5.5. Multilayer networks provide inside approximations

As well as changing the inner functions, there is the possibility of using multi-layer feedforward networks in $\vec{C}^n$. The impossibility results for inside approximation do *not* carry over.

Corollary 5.1.7. *Suppose that S_g^1 contains a continuous generalized sigmoid. Then for all $k \geq 2$, $S_{g,k}^n$ is d_∞-inside dense for $\vec{C}^n$.*

5.6. Limited weights in infinite dimensional contexts

With tighter restrictions on g or S_g^1, it is possible to loosen the restriction on $\mathcal{A}$ in Theorem 5.1. Suppose that g is analytic and not a polynomial, that $\mathcal{B}$ is convex, and that the span of $\mathcal{B}$ satisfies the conditions on $\mathcal{A}$ in Theorem 5.1. In this case, span($g\mathcal{B}$) satisfies the conclusions of Theorem 5.1. Being a convex set with a rich span is the infinite dimensional analog of being a limited set of weights in a finite dimensional setting.

[6] Funahashi and Nakamura (1993), Chen and Chen (1995b), Sandberg (1996), Sandberg and Xu (1996).

6. Previous literature and present novelty

Chen, Chen, and Liu's (CCL's) (Chen et al., 1995) Theorem 1 shows that if g is a generalized sigmoid, then S_g^n is d_∞-outside dense for $\vec{C}^1$ Theorems 2.1 and 2.1′ use the distinction between inside and outside approximation and expand CCL's Theorem 1 to characterizations of activation functions with approximation properties for functions on $\mathbb{R}^1$ with good limit behavior at infinity. These two results and the examples completely answer the questions and conjectures in CCL; Huang and Babri give a rather more involved proof of Theorem 2.1′.

CCL's (Chen et al., 1995) Theorem 2 shows that if S_g^1 is d_∞-outside dense for $\vec{C}^1$ then for all $n \geq 2$, S_g^2 is d_∞-outside dense for $\mathring{C}^n$. This is the difficult part of Theorem 3.1′. Theorem 3.1 provides a cautionary note, showing that it is not possible to use SLFF networks to approximate $\mathring{C}^n$ using networks that are themselves in $\mathring{C}^n$, $n \geq 2$. The results here show that inside and outside approximation are very different for $n \geq 2$. Two further explications of the difference come from Lemma B. First, Lemma B implies that close outside approximations to elements of $\mathring{C}^n$, $n \geq 2$, must have the property that no small subset of the approximating SLFF network has significant influence on the whole network. Second, Lemma B implies Theorem 3.2, which shows that, at best, only functions with no cross effects can be approximated from the inside using SLFF networks.

Corollary 5.1.7 provides a possibility result for inside approximation to contrast with Theorem 3.2 in the case of more than 2 dimensions. Corollary 5.1.7 shows that multi-layer networks are inside approximators, Theorem 3.2 shows that single layer networks are not.

The possibility of using limited weights for networks makes them more credible as models of natural phenomena, and makes them easier to implement. Theorem 3.3 shows that d_C-denseness with limited weights is possible for all non-polynomial analytic activation functions.[7] However, Theorem 3.4 shows that d_∞-outside denseness for classes of functions with good limit behavior at infinity is not possible using SLFF networks with limited weights.

It is possible to homeomorphically map the domains of the functions in $\vec{C}^n$ and $\mathring{C}^n$ to compact subsets of $\mathbb{R}^n$. Taking the composition of affine functions with the homeomorphism leads to classes of feedforward networks with non-linear functions as the inner functions, that is, as the input-to-hidden-layer connections. Theorem 5.1 and its many direct corollaries examine the implications of having general classes of inner functions for feedforward networks. These include the extension of approximation results to multi-layer networks, the approximation of operators on infinite dimensional spaces, and the (previously mentioned) possibility results for inside approximation using multi-layer networks.

[7] A strict superset of this class of activation functions without restrictions on the weights was studied in Leshno et al. (1993).

7. Proofs

Part of the first proof adapts Cybenko's (Cybenko, 1989) dual space approach to give a quick alternative proof of CCL's result that S_g^1 is d_∞-outside dense for $\vec{C}^1$ if g is a generalized sigmoid.

Proof of Theorem 2.1. Fix $\epsilon < 1/2$. If S_g^1 contains no continuous generalized sigmoids, then $S_g^1 \cap \vec{C}^1$ cannot uniformly approximate the logistic function to within ϵ.

If S_g^1 contains a continuous generalized sigmoid, then it contains a continuous generalized sigmoid, σ, with limits of 0 and 1 at negative and positive ∞ respectively. The d_∞-closure of $S_\sigma^1 \subset S_g^1$ is a proper subset of $\vec{C}^1$ if and only if there exists a non-zero ν in the dual space of $(\vec{C}^1, d_\infty)$, the set of countably additive, finite, signed Borel measures on $[-\infty, +\infty]$, that annihilates S_σ^1. Showing that $\int S_\sigma^1 \, d\nu = 0$ implies ν is zero completes the proof.

For each $r \in N = \{x \in \mathbb{R}^1 : \nu(\{x\}) = 0\}$, define uniformly bounded sequences of functions in S_σ^1, $h_r^\alpha(x) = \sigma(\alpha(x - r))$, and $\eta_r^\alpha(x) = \sigma(\alpha(r - x))$. These converge $|\nu|$-a.e. to $1_{(r,+\infty]}(x)$ and $1_{[-\infty,r]}(x)$ respectively. As $\int S_\sigma^1 \, d\nu = 0$, for all α, $r < s$, $r, s \in N$, $\int h_r^\alpha \, d\nu = \int (h_r^\alpha - h_s^\alpha) \, d\nu = \int \eta_s^\alpha \, d\nu = 0$. Dominated convergence implies that $\nu([-\infty, r]) = \nu((r, s]) = \nu((r, +\infty]) = 0$ for all $r, s \in N$. As N is dense, $\nu = 0$ on a field generating the Borel σ-field, so that $\nu = 0$.

Roughly speaking, Theorem 2.1′ follows from Theorem 2.1 and the triangle inequality.

Proof of Theorem 2.1′. If the d_∞-closure of S_g^1 contains $\vec{C}^1$, then it must contain the logistic function, a continuous, generalized sigmoid.

Now suppose that the d_∞-closure of S_g^1 contains a continuous generalized sigmoid, σ. Pick arbitrary target $h \in \vec{C}^1$ and arbitrary $\epsilon > 0$. It is sufficient to show that there is a function $f \in S_g^1$ such that $d_\infty(f(x), h(x)) < \varepsilon$. From Theorem 2.1, we know that there is a function $\varphi \in S_\sigma^1$ such that $d(\varphi(x), h(x)) < \varepsilon/2$. As $\vartheta \in S_\sigma^1$, it can be expressed as

$$\varphi(x) = \sum_{j=1}^{J} \beta_j \sigma(w_j x + b_j). \tag{13}$$

Pick positive $\delta < \epsilon/2JB$ where $B > \max_j |\beta_j|$. As σ is in the d_∞-closure of S_g^1, there is a function $\eta \in S_g^1$ such that $d_\infty(\eta(x), \sigma(x)) < \delta$. Set $f(x) = \sum_{j=1}^{J} \beta_j \eta(w_j x + b_j)$ so that $f \in S_g^1$. Finally,

$$\begin{aligned} d_\infty(f(x), h(x)) &\leq d_\infty(f(x), \varphi(x)) + d_\infty(\varphi(x), h(x)) \\ &< \varepsilon/2 + JB \cdot \varepsilon/2JB = \varepsilon. \end{aligned} \tag{14}$$

The following is somewhat more compact than Huang and Babri's (Huang and Babri, 1997) proof.

Proof of Theorem 2.2. If $g \in \mathring{M}_b^1$ (respectively $\mathring{C}^1$), then

F_g^1 is a subset of the d_∞-closed vector space $\mathring{M}_b^1$ (respectively $\mathring{C}^1$), implying that the d_∞-closure of span(F_g^1) is a subset of $\mathring{M}_b^1$ (respectively $\mathring{C}^1$).

Proof of Theorem 2.4. As $g \in \vec{M}^1$ the numbers $g_* = \lim_{x\to-\infty} g(x)$and $g^* = \lim_{x\to+\infty} g(x)$ exist. Fix an open subset $\mathcal{O}$ of $\mathbb{R}^2$ with the property that the first component is always the same sign, say positive. For any $(w_j, b_j)_{j=1}^J$ in $\mathbb{O}$, $\lim_{x\to-\infty} \sum_{j=1}^J \beta_j g_{(w_j,b_j)}(x)$ is equal to $g_*{\cdot}B$ and $\lim_{x\to+\infty} \sum_{j=1}^J \beta_j g_{(w_j,b_j)}(x)$ is equal to $g^*{\cdot}B$ where $B = \sum_{j=1}^J \beta_j$. In other words, the values of $\sum_{j=1}^J B_j g_{(w_j,b_j)}$ are linearly dependent at $\pm\infty$, showing that $S_g^1(\mathcal{O})$ cannot be d_∞- outside dense for $\vec{C}^1$.

Proof of Theorem 3.1. The proof will be given for $\mathring{C}^2$. The general case is parallel, but somewhat more notation intensive.

If the activation function g is constant, then the conclusion of the Theorem holds trivially. Suppose that $g \in \vec{C}^1$ is not constant.

Lemma A. *If $g \in \vec{C}^1$ is not constant, then $g_{(w,b)} \in \vec{C}^2$ if and only if w takes one of the following two forms*: $(w_1,0)$ *or* $(0,w_2)$.

Proof of Lemma A. Suppose first that $w = (w_1, 0)$ or $w = (0, w_2)$. If a sequence x^α converges in $[-\infty, +\infty]^2$, then $\lim_\alpha g(w{\cdot}x^\alpha + b)$ exists because $g \in \vec{C}^1$. This shows that $g_{(w,b)} \in \vec{C}^2$.

For the reverse implication, suppose that $w_1 \neq 0$ and $w_2 \neq 0$. It is sufficient to show that there are sequences $x^{i,\alpha}, i = 1, 2$, going to the same point in $[-\infty, +\infty]^2$ such that

$$\lim_\alpha g_{(w,b)}(\mathbf{x}^{1,\alpha}) \neq \lim_\alpha g_{(w,b)}(\mathbf{x}^{2,\alpha}). \tag{15}$$

Let ℓ^i be the line of (x_1,x_2) such that $w_1x_1 + w_2x_2 + b = r^i, i = 1, 2$ where $g(r^1) \neq g(r^2)$. These are parallel lines with equations

$$x_2 = -(w_1/w_2)x_1 + (r^i - b)/w_2 \tag{16}$$

where $(w_1/w_2) \neq 0$. Let $x_1^{i,\alpha} = \alpha, x_2^{i,\alpha} = -(w_1/w_2)x_1^{i,\alpha} + (r^i - b)/w_2$. The sequences $\mathbf{x}^{i,\alpha} = (x_1^{i,\alpha}, x_2^{i,\alpha})$ both go to $(+\infty, -\text{sgn}(w_1/w_2)\infty) \in [-\infty, +\infty]^2$ as $\alpha \to \infty$, yet $g_{(w,b)}(\mathbf{x}^{i,\alpha}) \equiv g(r_i)$, so the two limits are different, completing the proof of Lemma A.

From Lemma A we have the following.

Lemma B. *If $g \in \vec{C}^1$ is not constant and $f \in S_g^2 \cap \vec{C}^2$, then f can be expressed as a finite linear combination $f = \sum_{j=1}^J \beta_j g_{(w_j,b_j)}$ where each w_j is of the form given in Lemma A.*

Proof of Lemma B. Basic linear algebra implies that any $f \in S_g^2$ can be expressed as a finite linear combination of linearly independent $g_{(w_i,b_i)}$ for i in some finite set I. If for some $K \subset I, \#K \geq 2$, $\sum_{k\in K} \beta_k g_{(w_k,b_k)}$ is the constant function, replace the set of functions $\{g_{(w_k,b_k)} : k \in K\}$ by a single function $\beta g_{(0,b)}$. Relabel the new set of functions again as $g_{(w_j,b_j)}$, $j \in J$.

In the (perhaps) new, but still linearly independent set of functions, let K be the set of j such that w_j does not have the pattern in Lemma A. It is sufficient to show that K is empty. If K is not empty, then there is a non-empty $K' \subset K$, where K' is a maximal set of co-linear, non-zero $w_j, j \in K$. As the $g_{(wj,bj)}$ are linearly independent, the function $h = \sum_{j\in K'} \beta_j g_{(w_j,b_j)}$ is, by construction, non-constant. Pick r^1, r^2 such that $h(r^1) \neq h(r^2)$. As in Lemma A, it is possible to move to one of the corners of $[-\infty, +\infty]^2$ along sequences $\mathbf{x}^{i,\alpha}$ that stay on two parallel lines so that $h(\mathbf{x}^{i,\alpha}) \equiv h(r^i)$. Fix arbitrary $\delta, 0 < \delta < |h(r^1) - h(r^2)|$. For $\bar{\alpha}$ sufficiently large, the variation of $\sum_{j\notin K'} \beta_j g_{(w_j,b_j)}(\mathbf{x}^{i,\alpha})$ for $\alpha \geq \bar{\alpha}$ must be less than δ because $g \in \vec{C}^1$ and each $w_j{\cdot}x^{i,\alpha} + b_j$ is tending to either $+\infty$ or $-\infty$ when $j \notin K'$. But this contradicts $f \in \vec{C}^2$, showing that K' must be empty, implying that K must be empty, and completing the proof of Lemma B.

With Lemma B, the proof of Theorem 3.1 can be completed.

Pick arbitrary non-constant $h \in \mathring{C}^2$, let $r = \lim_{\|x^\alpha\|\to\infty} h(\mathbf{x}^\alpha)$, and let $\nu > 0$ be the variation of h, that is, the difference between the supremum and the infimum of h. The proof will be complete once we show that for $\varepsilon < \nu/6$, there is no $f \in S_g^2 \cap \vec{C}^2$ that is uniformly within ε of h. For the purposes of contradiction, suppose that $f \in S_g^2 \cap \vec{C}^2$ is uniformly with ε of h for some $\varepsilon < \nu/6$.

From Lemma B, we know that any $f \in S_g^2 \cap \vec{C}^2$ can be expressed as a finite linear combination $f = \sum_{i\in K} \beta_i g_{(w_i,b_i)}$ where each w_i is of the form given in Lemma A. Let K_1 be the set of $i \in K$ such that w_i is of the form $(w_{i,1},0)$, and let $K_2 = K\backslash K_1$. Let

$$G_1(x_1,x_2) = \sum_{i\in K_1} \beta_i g(w_{i,1}x_1 + b_i), \text{ and } G_2(x_1,x_2) = \sum_{i\in K_2} \beta_i g(w_{i,2}x_1 + b_i), \tag{17}$$

so that $f(x_1,x_2) = G_1(x_1,x_2) + G_2(x_1,x_2)$. Note that G_1 does not depend on x_2 and G_2 does not depend on x_1.

As $f \in \vec{C}^2$ and is uniformly within ε of $h \in \mathring{C}^2$, we know that for all x_2,

$$\lim_{\alpha\to\infty} f(\alpha,x_2) = \lim_{\alpha\to\infty} G_1(\alpha,x_2) + G_2(\alpha,x_2) \tag{18}$$

exists and is in the interval $[r - \varepsilon, r + \varepsilon]$. Therefore varying x_2 causes f to vary by at most 2ε. In a parallel fashion, varying x_1 causes f to vary of at most 2ε so that f varies by at most 4ε. But the variation of h is greater than 6ε. This means that the sup norm distance between h and f must be greater than ε, a contradiction that completes the proof.

Proof of Theorem 3.1′. If g is a generalized sigmoid, the sufficiency result in Theorem 2 of CCL (Chen et al., 1995). If g is not a generalized sigmoid but the d_∞ closure of S_g^1 contains one, then apply the triangle inequality as in the proof of Theorem 2.1′.

The failure of necessity can be had by taking $g \in \mathring{C}^1$ to be e.g. the density of a Gaussian random variable. In this case, Theorem 2.2 implies that the d_∞ closure of S_g^1 cannot contain a generalized sigmoid. However, let ν be Lebesgue measure on S^{n-1}, the surface of the unit ball in $\mathbb{R}^n$. The function

$$f_{\lambda,c}(\mathbf{x}) = \int g(\lambda w\cdot(x - c))d\nu(w) \tag{19}$$

is non-negative, has strictly positive integral, and has the property that $\|\mathbf{x}\| \to \infty$ implies $f_{\lambda,c}(\mathbf{x}) \to 0$. By the uniform strong law of large numbers, taking the average of a sufficiently large, i.i.d. random sample of w_j distributed according to ν gives a sequence of elements of S_g^n converging in the metric d_∞ to $f_{\lambda,c}$. The functions $f_{\lambda,c}$ are spherically symmetric, hence the span of the $f_{\lambda,c}$ contains a radial basis network. It is easy to show that this network is d_∞-dense in $\mathring{C}^n$. To complete the proof, apply the triangle inequality again as in the proof of Theorem 2.1′.

Proof of Theorem 3.2. Immediate from Theorem 2.1 and Lemma B.

Proof of Theorem 2.3 and 3.3. If g is a polynomial of degree k, then $S_g^n(\mathcal{O})$ is a subset of the polynomials of degree k for any $\mathcal{O}$. Polynomials of degree k cannot approximate polynomials of degree $k + 1$ or higher over compact sets with non-empty interior.

If g is analytic and not a polynomial, then in particular it is not a polynomial on some interval. Using Hornik (1993) or Leshno et al. (1993), this implies that $S_g^n(\mathbb{R}^{n+1})$ is d_C-dense for C^n. The dual space for (C^n,d_C) is the set of compactly supported, finite, signed, countably additive Borel measures of $\mathbb{R}^n$. Pick arbitrary $\mathcal{O}$ with non-empty interior and consider the function $\psi(w,b) = \int g_{(w,b)}(\mathbf{x})\,\mathrm{d}\nu(\mathbf{x})$ where ν is in the dual space of C^n. It is sufficient to show that if ψ is equal to 0 on $\mathcal{O}$, then ν must be 0. As ψ is analytic, it is equal to 0 on $\mathcal{O}$ if and only if it is equal to 0 on all of $\mathbb{R}^{n+1}$. But $S_g^n(\mathbb{R}^{n+1})$ is d_C-dense, implying that ν must be 0.

Proof of Theorem 3.4. Pick $h \in \mathring{C}^n$, $n \geq 2$, with $\lim_{\|\mathbf{x}\|\to\infty} h(\mathbf{x}) = 1$. Pick $\mathcal{O}$ to be an open subset of strictly positive vectors in $\mathbb{R}^{n+1}$. All feedforward networks $f \in S_g^n(\mathcal{O})$ can be expressed as integral transformations of the form $f_\nu(\mathbf{x}) = \int_{\mathcal{O}} g_{(w,b)}(\mathbf{x})\,\mathrm{d}\nu(w,b)$ where ν is a finitely supported measure on $\mathcal{O}$. For $\mathbf{x}$ with a single positive infinite component, $f_\nu(\mathbf{x})$ is equal to $\nu(\mathcal{O})g_*$, with a single negative infinite component, $\nu(\mathcal{O})g_*$ where $g^* = \lim_{x\to+\infty} g(x)$ and $g_* = \lim_{x\to-\infty} g(x)$. If $d_\infty(f_\nu, h) < \varepsilon$, then $|\nu(\mathcal{O})g_* - 1| < \varepsilon$ and $|\nu(\mathcal{O})g^* - 1| < \varepsilon$. If S_g^n is $d\infty$-outside dense for $\mathring{C}^n$, there must be numbers $\nu(\mathcal{O})$ making these inequalities true for all $\epsilon > 0$. This implies that $g_* \neq g^*$ is not possible. This in turn implies that g is not a generalized sigmoid.

Proof of Theorem 5.1. To prove (a), note that the class of functions $\mathrm{span}(\cos\circ\mathscr{A})$ is an algebra of continuous functions that separates points, and its ρ_K-closure contains the constant functions. Therefore it is ρ_K-inside dense for $C(K)$. Pick arbitrary $\epsilon > 0$ and $\eta(x) = \sum_{j=1}^J \beta_j\cos(a_j(x)) \in \mathit{span}(\cos\circ\mathscr{A})$. It is sufficient to show that there is a function $h \in \mathrm{span}(g\circ\mathscr{A})$ such that $\rho_K(h,g) < \epsilon$. Let $K' \subset \mathbb{R}^1$ be a compact set containing $\bigcup_{j=1}^J a_j(K)$. By assumption, it is possible to find $f \in S_g^1$ that is within $\rho_{K'}$-distance $\delta > 0$ of cos where $\delta < \epsilon/(J\cdot B)B > \max_j|\beta_j|$. Setting $h = \sum_{j=1}^J \beta_j f(a_j(x))$ completes the proof of (a). The proof of (b) is a direct parallel.

Proof of Corollary 5.1.1. Under the assumptions given, S_g^n and $S_{g,k}^n$ satisfy the conditions on $\mathscr{A}$ in Theorem 5.1

Proof of Corollary 5.1.2. The class of functions $l_j(x) + b_j$ satisfy the conditions on $\mathscr{A}$ used in Theorem 5.1.

Proof of Corollary 5.1.3. Pointwise convergence of the continuous functions in the weak*-dense subset of X^* to a continuous function in X^* is uniform over compact sets.

Proof of Corollary 5.1.4. The proof of Corollary 5.1.1 applies.

Proof of Corollaries 5.1.5 and 5.1.6. In both case, the class of functions $\mathscr{A}\circ\varphi$ satisfy the conditions of Theorem 5.1.

Proof of Corollary 5.1.7. By Theorem 3.2, $S_g^n \cap \vec{C}^n$ is d_∞-inside dense for $\mathscr{C}$. The class of functions $\mathscr{C}$ satisfies the conditions on $\mathscr{A}$ used in Theorem 5.1, implying that $S_g^n \cap \vec{C}^n$ also satisfies them.

References

Cardaliaguet, P., & Euvrard, G. (1992). Approximation of a function and its derivatives with a neural network. *Neural Networks, 5* (2), 207–220.

Chen, T. -P., & Chen, H. (1995a). Approximation capability to functions of several variables, Nonlinear functionals, and operators by radial basis function neural networks. *IEEE Transactions on Neural Networks, 6* (4), 904–910.

Chen, T. -P., & Chen, H. (1995b). Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Transactions on Neural Networks, 6* (4), 911–917.

Chen, T.-P., Chen, H. & Liu, R.-W. (1990). A constructive proof of Cybenko's approximation theorem and its extensions, 163–168 in computing science and statistics, LePage and Page (eds.). Proceedings of the 22nd Symposium on the Interface. New York: Springer-Verlag.

Chen, T. -P, Chen, H., & Liu, R. -W. (1995). Approximation capability in $C(;rcam\mathbb{R}^n)$ by multilayer feedforward networks and related problems. *IEEE Transactions on Neural Networks, 6* (1), 25–30.

Chui, C. K., & Li, X. (1992). Approximation by ridge functions and neural networks with one hidden layer. *Journal of Approximation Theory*, *70*, 131–141.

Chui, C. K., Li, X., & Mhaskar, H. N. (1996). Limitations of the approximation capabilities of neural networks with one hidden layer. *Advances in Computational Mathematics*, *5* (2–3), 233–243.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function, Mathematics of Control. *Signals and Systems*, *2*, 303–314.

Elanayar, S., & Shin, Y. C. (1994). Radial basis function neural network for approximation and estimation of nonlinear stochastic dynamic systems. *IEEE Transactions on Neural Networks*, *5* (4), 594–603.

Funahashi, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, *2*, 183–192.

Funahashi, K., & Nakamura, Y. (1993). Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Networks*, *6* (6), 801–806.

Gallant, A.R., White, H. (1988). There exists a neural network that does not make avoidable mistakes. Proceedings of the Second International Joint Conference on Neural Networks, I 593-606. San Diego: SOS Printing.

Guillemin, V., & Pollack, A. (1974). Differential Topology. Englewood Cliffs: Prentice Hall.

Hecht-Nielson, R. (1989). Theory of the back propagation neural network. Proceedings of the International Joint Conference on neural networks, I, 593-606. San Diego: SOS Printing.

Hornik, K. (1991). Approximation Capabilities of multilayer feedforward networks. *Neural Networks*, *4* (2), 251–257.

Hornik, K. (1993). Some new results on neural network approximation. *Neural Networks*, *6* (8), 1069–1072.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*, 359–366.

Hornik, K., Stinchcombe, M., & White, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, *3*, 551–560.

Huang, G.-B. & Babri, H. (1997). On approximation capability in $C(\bar{R}^n)$ by multilayer feedforward networks and related problems, photocopy, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

Hurd, A. E., & Leob, P. A. (1985). An introduction to nonstandard real analysis. Orlando: Academic Press.

Ito, Y. (1993). Approximation capability of layered networks with sigmoid units on tow layers. *Neural Computation*, *6* (6), 1233–1243.

Kelley, J. L. (1955). General topology. New York: Springer-Verlag.

Leshno, M., Lin, V. Y., Pinkus, A., & Shocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, *6* (6), 861–867.

Li, X. (1996). Simultaneous approximation of multivariate functions and their derivatives by neural networks with one hidden layer. *Neurocomputing*, *12* (4), 327–343.

Mhaskar, H. N. (1993). Approximation properties of a multilayered feedforward artificial neural network. *Advances in Computational Mathematics*, *1* (1), 61–80.

Mhaskar, H. N., & Micchelli, C. A. (1992). Approximation by superposition of sigmoidal and radial basis functions. *Advances in Applied Mathematics*, *13* (3), 350–373.

Mhaskar, H. N., & Micchelli, C. A. (1995). Degree of approximation by neural and translation networks with a single hidden layer. *Advances in Applied Mathematics*, *16* (2), 151–183.

Park, J. Y., & Sandberg, I. W. (1993). Approximation and radial-basis-function networks. *Neural Computation*, *5* (2), 305–316.

Park, J. Y., & Sandberg, I. W. (1994). Nonlinear approximations using elliptic basis function networks. *Circuits, Systems and Signal Processing*, *13* (1), 99–113.

Sandberg, I. W. (1996). Notes on weighted norms and network approximation of functionals. *IEEE Transactions on Circuits and Systems I – Fundamental Theory and Applications*, *43* (7), 600–601.

Sandberg, I. W., & Xu, L. (1996). Network approximation of input–output maps and functionals. *Circuits Systems and Signal Processing*, *15* (6), 711–725.

Stinchcombe, M. (1995). Precision and approximate flatness in artificial neural networks. *Neural Computation*, *7* (5), 1021–1039.

Stinchcombe, M. & White, H. (1989). Universal approximation using feedforward networks with non-sigmoid hidden layer activation functions, Proceedings of the International Joint Conference on Neural Networks, I, 613-617. San Diego: SOS Printing.

Stinchcombe, M. & White, H. (1990). Approximating and learning unknown mappings using multilayer feedforward networks with bounded weights, Proceedings of the International Joint Conference on Neural Networks, Washington, D.C. III, 7-16. San Diego: SOS Printing (1990).

Stinchcombe, M. & White, H. (1992). Using feedforward networks to distinguish multivariate populations, in Proceedings of the International Joint Conference on Neural Networks, IEEE Press, New York, I: 788-793.