An elementary proof of a universal approximation theorem

Chris Monico

Department of Mathematics and Statistics
Texas Tech University
e-mail: c.monico@ttu.edu

Abstract

In this short note, we give an elementary proof of a universal approximation theorem for neural networks with three hidden layers and increasing, continuous, bounded activation function. The result is weaker than the best known results, but the proof is elementary in the sense that no machinery beyond undergraduate analysis is used.

1 Introduction

There are several versions of *universal approximation theorems* known, including the very well-known ones from [1, 2, 3]. Each of them states that some collection of neural networks is dense in some space of continuous functions with respect to the uniform norm. In this short note, we present what we believe to be a new and atypically elementary proof of one such theorem. If σ is a 0-1 squashing function (a.k.a. a *sigmoidal function*), we show that the collection of neural networks with three hidden layers and activation function σ (except at the output) is dense in the space $C(K)$ of real-valued continuous functions on a compact set $K \subset \mathbb{R}^n$. The result given here is weaker than the best known results, but the argument relies only on basic results about compact sets and continuous functions which are generally covered in an undergraduate analysis course; it is really nothing more than an exercise in "epsilon chasing," though the underlying intuitive motivation is fairly natural.

2 Notations

For the entirety of this note, we assume the following:

1. $\sigma : \mathbb{R} \longrightarrow \mathbb{R}$ is increasing, continuous, and $\lim_{x \to -\infty} \sigma(x) = 0$ and $\lim_{x \to \infty} \sigma(x) = 1$.
2. K is a compact subset of $\mathbb{R}^n$.

For example, the sigmoid function $\sigma(x) = 1/(1 + e^{-x})$ satisfies the first assumption. A function satisfying that assumption will be called a 0-1 *squashing function*. There's nothing particularly special about 0 and 1 here. The crucial properties used in our arguments are really just that σ is increasing, continuous, bounded, and non-constant. But the discussion is simplified with these additional limiting assumptions.

The compact set K will represent the domain of a function to be approximated. In particular, the unit cube $K = [0,1]^n$ is a special case.

Of course, it's not at all necessary for a neural network to use the same activation function in each hidden layer, but that will suffice for our purposes. Fix a positive integer n and let

$$\begin{aligned}\mathcal{N}_1 &= \{f \in C(K) \ : \ f(x_1,\dots,x_n) = a_0 + a_1x_1 + \cdots + a_nx_n, \\ &\qquad \text{for some } a_0,\dots,a_n \in \mathbb{R}\}, \\ \mathcal{N}_1^\sigma &= \{F \in C(K) \ : \ F = \sigma \circ f, \quad \text{for some } f \in \mathcal{N}_1\}.\end{aligned}$$

That is, $\mathcal{N}_1$ is the set of all affine functions of $x_1,\dots,x_n$. And $\mathcal{N}_1^\sigma$ is the set of all possible node output functions in Layer 1. For $k \geq 1$, we define

$$\begin{aligned}\mathcal{N}_{k+1} &= \{g \in C(K) \ : \ g = a_0 + a_1F_1 + \cdots + a_mF_m, \\ &\qquad \text{for some } F_1,\dots,F_m \in \mathcal{N}_k^\sigma, \ a_0,\dots,a_m \in \mathbb{R}\}, \\ \mathcal{N}_{k+1}^\sigma &= \{G \in C(K) \ : \ G = \sigma \circ g, \quad \text{for some } g \in \mathcal{N}_{k+1}\}.\end{aligned}$$

Thus, $\mathcal{N}_{k+1}^\sigma$ is the set of all possible node output functions in Layer $k+1$. It follows easily from the definitions above that $\mathcal{N}_k^\sigma \subset \mathcal{N}_{k+1}$, and if $g_1, g_2 \in \mathcal{N}_k$ then $(a_0 + a_1g_1 + a_2g_2) \in \mathcal{N}_k$ for all $a_0, a_1, a_2 \in \mathbb{R}$. We will use these facts frequently in the sequel.

3 Three separation lemmas

Roughly speaking, the strategy of the proof is to first show that σ separates points in $\mathbb{R}^n$ in a strong sense. We then show that $\mathcal{N}_2$ separates points from closed sets in a similarly strong sense, and that $\mathcal{N}_3$ separates closed sets from closed sets again in a strong sense. Finally, the theorem will be proven using this result about $\mathcal{N}_3$.

Lemma 3.1 *Let x_0 and x_1 be distinct real numbers. For each $\epsilon > 0$ there exist $s, t \in \mathbb{R}$ such that $\sigma(s + tx_0) < \epsilon$ and $\sigma(s + tx_1) > 1 - \epsilon$. If, in addition, $x_0 < x_1$ and $\epsilon < 1/2$, then $\sigma(s + tx) < \epsilon$ on the interval $(-\infty, x_0]$ and $\sigma(s + tx) > 1 - \epsilon$ on the interval $[x_1, \infty)$.*

Proof: Without loss of generality, we may assume $\epsilon < 1$. By definition of a 0-1 squashing function, there exist $y_0, y_1 \in \mathbb{R}$ such that $\sigma(y_0) = \epsilon/2$ and $\sigma(y_1) = 1 - \epsilon/2$. Since $x_1 - x_0 \neq 0$, the linear system

$$\begin{pmatrix} 1 & x_0 \\ 1 & x_1 \end{pmatrix} \begin{pmatrix} s \\ t \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \end{pmatrix},$$

has a solution, say $(s_0, t_0)^T$. Thus,

$$\begin{aligned}\sigma(s_0 + t_0x_0) &= \sigma(y_0) = \epsilon/2 < \epsilon, \quad \text{and} \\ \sigma(s_0 + t_0x_1) &= \sigma(y_1) = 1 - \epsilon/2 > 1 - \epsilon.\end{aligned}$$

Suppose, in addition, that $x_0 < x_1$ and $\epsilon < 1/2$. Since $\sigma(s+tx)$ is monotone and $\sigma(s+tx_0) < \epsilon < 1/2 < 1-\epsilon < \sigma(s+tx_1)$, it follows that $\sigma(s+tx)$ is increasing, which proves the remainder of the lemma. $\square$

The next lemma asserts that we can separate points from closed sets using functions in $\mathcal{N}_2$ in a strong sense. That is, one hidden layer suffices to separate points from closed sets in the described sense.

Lemma 3.2 *Let $B \subset K$ be a closed set, and $\boldsymbol{x}_0 \in K - B$. For each $\epsilon > 0$ there exists $g \in \mathcal{N}_2$ such that $g > 1-\epsilon$ on B and $g(\boldsymbol{x}_0) < \epsilon$.*

Proof: Without loss of generality, assume $0 < \epsilon < 1/3$. Let $\boldsymbol{b} \in B$. Since $\boldsymbol{b} \neq \boldsymbol{x}_0$, there exists $f_{\boldsymbol{b}} \in \mathcal{N}_1$ such that $f_{\boldsymbol{b}}(\boldsymbol{x}_0) < \epsilon/2$ and $f_{\boldsymbol{b}}(\boldsymbol{b}) > 1-\epsilon/2$. Let

$$U_{\boldsymbol{b}} = \{\boldsymbol{x} \in K \ : \ f_{\boldsymbol{b}}(\boldsymbol{x}) > 1-\epsilon\}.$$

Since $f_{\boldsymbol{b}}$ is continuous, $U_{\boldsymbol{b}}$ is open and $\boldsymbol{b} \in U_{\boldsymbol{b}}$, so $\{U_{\boldsymbol{b}}\}_{\boldsymbol{b}\in B}$ is an open cover of the compact set B. Let $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N \in B$ such that $\{U_{\boldsymbol{b}_1}, \ldots, U_{\boldsymbol{b}_N}\}$ is a cover of B.

By Lemma 3.1, there exist $s, t \in \mathbb{R}$ such that $\sigma(s+tx) < \epsilon/N$ on $(-\infty, \epsilon)$ and $\sigma(s+tx) > 1-\epsilon$ on $(1-\epsilon, \infty)$. For each $1 \leq j \leq N$, let $F_j = \sigma(s+tf_{\boldsymbol{b}_j})$. Then $F_j \in \mathcal{N}_1^\sigma$ and $F_j(\boldsymbol{x}_0) < \epsilon/N$ and $F_j \geq 1-\epsilon$ on $U_{\boldsymbol{b}_j}$.

Let $g = \sum_{j=1}^N F_j$, and it follows that $g \in \mathcal{N}_2$, $g(\boldsymbol{x}_0) < \epsilon$ and $g \geq 1-\epsilon$ on B. $\square$

The following lemma is the major tool in the proof of the universal approximation theorem we will present. It asserts that two hidden layers suffice to separate disjoint closed sets, in a sense very similar to the previous lemma. Its proof is quite similar to the previous one.

Lemma 3.3 *Let A and B be disjoint closed subsets of K. Then for each $\epsilon > 0$,*

(i) there exists $h \in \mathcal{N}_3$ such that $h < \epsilon$ on B and $h > 1-\epsilon$ on A,

(ii) there exists $H \in \mathcal{N}_3^\sigma$ such that $0 \leq H < \epsilon$ on B and $1-\epsilon < H \leq 1$ on A.

Proof: Again without loss of generality we assume $\epsilon \in (0, 1/3)$. For each $\boldsymbol{a} \in A$, by Lemma 3.2, there exists $\widetilde{g}_{\boldsymbol{a}} \in \mathcal{N}_2$ such that $\widetilde{g}_{\boldsymbol{a}} > 1-\epsilon/2$ on B and $\widetilde{g}_{\boldsymbol{a}}(\boldsymbol{a}) < \epsilon/2$. Let $g_{\boldsymbol{a}} = 1 - \widetilde{g}_{\boldsymbol{a}}$, so that $g_{\boldsymbol{a}} \in \mathcal{N}_2$, and $g_{\boldsymbol{a}} < \epsilon/2$ on B, and $g_{\boldsymbol{a}}(\boldsymbol{a}) > 1-\epsilon/2$. Let

$$U_{\boldsymbol{a}} = \{\boldsymbol{x} \in K \ : \ g_{\boldsymbol{a}}(\boldsymbol{x}) > 1-\epsilon\}.$$

Since $g_{\boldsymbol{a}}$ is continuous, each $U_{\boldsymbol{a}}$ is open. And since $\boldsymbol{a} \in U_{\boldsymbol{a}}$, it follows that $\{U_{\boldsymbol{a}}\}_{\boldsymbol{a}\in A}$ is an open cover of the compact set A. Let $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N \in A$ for which $\{U_{\boldsymbol{a}_1}, \ldots, U_{\boldsymbol{a}_N}\}$ is a cover of A.

Let $s, t \in \mathbb{R}$ for which $\sigma(s+tx) < \epsilon/N$ on $(-\infty, \epsilon)$ and $\sigma(s+tx) > 1-\epsilon$ on $(1-\epsilon, \infty)$. Set

$$h = \sum_{j=1}^N \sigma(s+tg_{\boldsymbol{a}_j}).$$

First note that $h \in \mathcal{N}_3$. If $\boldsymbol{a} \in A$, then $\boldsymbol{a} \in U_{\boldsymbol{a}_k}$ for some k, which implies that $g_{\boldsymbol{a}_k}(\boldsymbol{a}) > 1-\epsilon$ and so $h(\boldsymbol{a}) > 1-\epsilon$. And if $\boldsymbol{b} \in B$ it follows that $g_{\boldsymbol{a}_j}(\boldsymbol{b}) < \epsilon/2$ for all j, which gives $\sigma(s + g_{\boldsymbol{a}_j}(\boldsymbol{b})) < \epsilon/N$ for all j and hence $h(\boldsymbol{b}) < \epsilon$. This proves the first part.

Since σ is increasing, by Lemma 3.1 there exist $s, t \in \mathbb{R}$ such that $\sigma(s+tx) < \epsilon$ on the interval $(-\infty, \epsilon)$ and $\sigma(s+tx) > 1-\epsilon$ on the interval $(1-\epsilon, \infty)$. Then $H = \sigma(s+th)$ has the required properties.

$\square$

Recall that if K is a compact subset of $\mathbb{R}^n$ and $f : K \longrightarrow \mathbb{R}$ is continuous, then the *sup norm* of f, denoted $\|f\|_u$, is

$$\|f\|_u = \sup\{|f(\boldsymbol{x})| \;:\; \boldsymbol{x} \in K\}.$$

Theorem 3.4 (A Universal Approximation Theorem) *Let σ be a 0-1 squashing function, and $\mathcal{N}_k, \mathcal{N}_k^\sigma$ as previously defined. Let $T : K \longrightarrow \mathbb{R}$ be a continuous function. For each $\epsilon > 0$ there exists $f \in \mathcal{N}_4$ such that $\|f - T\|_u < \epsilon$; that is, $\mathcal{N}_4$ is dense in $C(K)$ with respect to the* sup *norm.*

Proof: By way of contradiction, suppose that $T : K \longrightarrow \mathbb{R}$ is a continuous function and

$$\inf_{f \in \mathcal{N}_4} \|f - T\|_u = \alpha > 0.$$

Let $\widehat{f} \in \mathcal{N}_4$ with $\alpha \leq \|\widehat{f} - T\|_u < 4\alpha/3$. Define

$$\begin{aligned} U^+ &= \left\{ \boldsymbol{x} \in K \;:\; \frac{\alpha}{3} \leq (\widehat{f} - T)(\boldsymbol{x}) \leq \frac{4\alpha}{3} \right\}, \\ U^- &= \left\{ \boldsymbol{x} \in K \;:\; -\frac{4\alpha}{3} \leq (\widehat{f} - T)(\boldsymbol{x}) \leq -\frac{\alpha}{3} \right\}. \end{aligned}$$

Since U^+, U^- are closed and disjoint, by Lemma 3.3 there exists $H \in \mathcal{N}_3^\sigma$ such that

$$0 \leq H < \frac{1}{6} \quad \text{on } U^-, \qquad \text{and } \frac{5}{6} < H \leq 1 \quad \text{on } U^+.$$

Consider $f = \widehat{f} - \alpha H + \frac{\alpha}{2} \in \mathcal{N}_4$. We claim that $\|f - T\|_u < \alpha$.

Suppose first that $\boldsymbol{x} \in U^+$. Then

$$\begin{aligned} (f - T)(\boldsymbol{x}) &= (\widehat{f} - T)(\boldsymbol{x}) - \alpha H(\boldsymbol{x}) + \frac{\alpha}{2} \\ &< \frac{4\alpha}{3} - \frac{5\alpha}{6} + \frac{\alpha}{2} = \alpha. \end{aligned}$$

We also have that

$$\begin{aligned} (f - T)(\boldsymbol{x}) &= (\widehat{f} - T)(\boldsymbol{x}) - \alpha H(\boldsymbol{x}) + \frac{\alpha}{2} \\ &\geq \frac{\alpha}{3} - \alpha + \frac{\alpha}{2} = -\frac{\alpha}{6} > -\alpha, \end{aligned}$$

And therefore $|f - T| < \alpha$ on U^+. Suppose now that $\boldsymbol{x} \in U^-$. Then

$$\begin{aligned}(f - T)(\boldsymbol{x}) &= (\widehat{f} - T)(\boldsymbol{x}) - \alpha H(\boldsymbol{x}) + \frac{\alpha}{2} \\ &\leq -\frac{\alpha}{3} + \frac{\alpha}{2} = \frac{\alpha}{6} < \alpha,\end{aligned}$$

and

$$\begin{aligned}(f - T)(\boldsymbol{x}) &= (\widehat{f} - T)(\boldsymbol{x}) - \alpha H(\boldsymbol{x}) + \frac{\alpha}{2} \\ &> -\frac{4\alpha}{3} - \frac{\alpha}{6} + \frac{\alpha}{2} = -\alpha,\end{aligned}$$

and hence $|f - T| < \alpha$ on U^-. Finally, suppose that $\boldsymbol{x} \in K - (U^- \cup U^+)$. Since $-\frac{4\alpha}{3} < (\widehat{f} - T)(\boldsymbol{y}) < \frac{4\alpha}{3}$ for all $\boldsymbol{y} \in K$, it follows that $-\frac{\alpha}{3} < (\widehat{f} - T)(\boldsymbol{x}) < \frac{\alpha}{3}$. From this and $0 \leq H \leq 1$ it follows that

$$\begin{aligned}(f - T)(\boldsymbol{x}) &= (\widehat{f} - T)(\boldsymbol{x}) - \alpha H(\boldsymbol{x}) + \frac{\alpha}{2} \\ &< \frac{\alpha}{3} + \frac{\alpha}{2} = \frac{5\alpha}{6} < \alpha,\end{aligned}$$

and

$$\begin{aligned}(f - T)(\boldsymbol{x}) &= (\widehat{f} - T)(\boldsymbol{x}) - \alpha H(\boldsymbol{x}) + \frac{\alpha}{2} \\ &> -\frac{\alpha}{3} - \alpha + \frac{\alpha}{2} = -\frac{5\alpha}{6} > -\alpha.\end{aligned}$$

But $f \in \mathcal{N}_4$ and $\|f - T\|_u < \alpha$ is a contradiction, and so the theorem is proven. $\square$

4 Remarks

There are a number of easily deduced corollaries that extend the result. For example, if σ is strictly increasing, we can obtain a similar result using $\mathcal{N}_4^\sigma$ for continuous functions from K to the image $\sigma(\mathbb{R})$. Of course another corollary is that functions $C(K, \mathbb{R}^m)$ can be uniformly approximated by functions in $(\mathcal{N}_4)^m$, and these two extensions could also be combined. But these are all very standard and obvious extensions which apply quite generally.

There are several places where this argument is 'wasteful'; that is, places where we are not using the full power of a three hidden layer neural network. This leaves room for the possibility that a very similar technique might be used to deduce the same conclusion for neural networks with two hidden layers.

We are grateful to Anupam Pal Choudhury for carefully reading the first draft of this note and pointing out several places where the argument was unclear.

References

[1] George V. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2:303–314, 1989.

[2] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[3] Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.