# ON APPROXIMATIONS OF FUNCTIONS BY DEPTH-TWO NEURAL NETWORKS

Santosh S. Venkatesh[1]

Department of Electrical Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA

*Abstract* — **The simple Pythagorean notion of orthogonal projections is used to show that depth-two sigmoidal neural networks can approximate any square-integrable function with compact support in $\mathbb{R}^n$ with arbitrarily small integrated squared-error.**

## I. Introduction

A recent result of G. Cybenko [1] asserts that, under mild conditions on the univariate function $\sigma$, linear combinations of the form

$$\widehat{f}(\mathbf{x}) = \sum_{i=1}^{N} c_i \sigma\big(\mathbf{w}_i^T \mathbf{x} - w_{i0}\big) \tag{1}$$

can uniformly approximate any continuous function $f(\mathbf{x})$ with support in the unit hypercube of $\mathbb{R}^n$.[2] The form $\sigma\big(\mathbf{w}_i^T\mathbf{x} - w_{i0}\big)$ can be interpreted as the response of a neuron with weight vector $\mathbf{w}_i \in \mathbb{R}^n$ and real threshold $w_{i0}$ to an input $\mathbf{x} \in \mathbb{R}^n$; the representation (1) hence corresponds to a depth-two neural network (or a neural network with one hidden layer). Cybenko's result can hence be succintly, if loosely, stated as follows: *Depth-two neural networks are universal approximators.* The purpose of this note is to show that the flavour of Cybenko's result is accessible by means of a direct elementary argument using the Pythagorean notion of orthogonal projections.[3]

## II. Function Approximation

Using Cybenko's terminology, we restrict attention to *discriminatory* functions $\sigma$ which are continuous and such that, for constants $a \neq b$, $\sigma(z) \to a$ as $z \to -\infty$ and $\sigma(z) \to b$ as $z \to \infty$. Let $\mathcal{C}$ be the inner-product space of continuous real-valued functions $f(\mathbf{x})$ defined on the unit hypercube $\mathbb{E}^n$ with the natural inner-product $\langle f, g\rangle = \int f(\mathbf{x})g(\mathbf{x})\,dx$ and the induced $L^2$-norm $\|f\| = \sqrt{\langle f, f\rangle}$. Denote the Hilbert space of square-integrable functions obtained as the $L^2$-closure of the set $\mathcal{C}$ by $\mathcal{L}^2$. Finally, let $\mathcal{N}_2^\sigma$ denote all functions $\widehat{f}(\mathbf{x})$ with support in $\mathbb{E}^n$ which can be obtained as a finite linear combination of the form (1).

**Theorem 1** *For any discriminatory $\sigma$, the family $\mathcal{N}_2^\sigma$ is dense in $\mathcal{L}^2$. In particular, for any $f \in \mathcal{C}$ and $\epsilon > 0$, there exists a finite linear combination of the form*

$$\widehat{f}(\mathbf{x}) = \sum_{i=1}^{N} c_i \sigma\big(\mathbf{w}_i^T \mathbf{x} - w_{i0}\big),$$

*for which*

$$\|f - \widehat{f}\| = \sqrt{\int \big|f(\mathbf{x}) - \widehat{f}(\mathbf{x})\big|^2 d\mathbf{x}} < \epsilon.$$

Proof: Consider the closure $\mathrm{cl}\,\mathcal{N}_2^\sigma$ of the set $\mathcal{N}_2^\sigma$. Clearly, $\mathrm{cl}\,\mathcal{N}_2^\sigma$ is a closed subspace of $\mathcal{L}^2$. Suppose that $\mathrm{cl}\,\mathcal{N}_2^\sigma \neq \mathcal{L}^2$. By the orthogonal projection theorem, we can then conclude that there exists a nonzero $h \in \mathcal{L}^2$ orthogonal to the closed subspace $\mathrm{cl}\,\mathcal{N}_2^\sigma$, viz., $\langle g, h\rangle = 0$ for every $g \in \mathrm{cl}\,\mathcal{N}_2^\sigma$.

Now consider any function of the form $g(\mathbf{x}) = \sigma\big(\mathbf{w}^T\mathbf{x} - w_0\big)$. Clearly, $g \in \mathrm{cl}\,\mathcal{N}_2^\sigma$. It follows that

$$0 = \langle g, h\rangle = \int \sigma\big(\mathbf{w}^T\mathbf{x} - w_0\big) h(\mathbf{x})\,d\mathbf{x} \tag{2}$$

for all $\mathbf{w} \in \mathbb{R}^n$ and $w_0 \in \mathbb{R}$. Now, the family of functions $\sigma_\lambda(z) = \sigma(\lambda z)$ converge boundedly and pointwise to a step function as $\lambda \to \infty$. A direct argument then shows (or invoke the Lebesgue bounded convergence theorem) that $\int h(\mathbf{x})\,d\mathbf{x} = 0$ over all planes $\mathbf{w}^T\mathbf{x} = w_0$ and half-spaces $\mathbf{w}^T\mathbf{x} > w_0$. As these planes and half-spaces generate all Borel sets, we must have $h = 0$, contradicting the hypothesis. It follows that the subspace $\mathcal{N}_2^\sigma$ is dense in $\mathcal{L}^2$. ∎

Remark: Cybenko [1] establishes (2) by appeal to the Hahn-Banach theorem and the Riesz representation theorem. The $L^2$-metric used here forfends the added technical apparatus needed to show uniform pointwise approximation.

## References

[1] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Math. Control, Signals, Systems*, vol. 2, pp. 183–192, 1989.

[2] K. Hornik, M. Stitchcomb, and H. White, "Multilayer neural networks are universal approximators," *Neural Networks*, vol. 2, pp. 359–368, 1989.

[1] This work was supported by the Air Force Office of Scientific Research under grant F49620-93-1-0120.

[2] Many variations on this theme exist in the literature; cf. Hornik, *et al.* [2], for instance.

[3] With historical antecedents ...
"About binomial theorem I'm teeming with a lot o' news; With many cheerful facts about the square of the hypotenuse." — W. S. Gilbert, *The Pirates of Penzance.*

0 - 7803-2015-8/94/$4.00 ©1994 IEEE