Universal Approximation Theorem 1

Taeyoung Lee

July 23, 2022

Abstract

Don't be afraid of reading papers with mathematics!
Are you sick of papers that mention a bunch of mathematical theorems you have never heard of and searching for them? Or figuring out what on this ridiculous mathematical world is going on? Don't worry. Leave the 'searching for mathematical this and that' part on me. Now you only have to google for other less boring stuffs! We dive together into papers and crack all the boring mathematical parts. By the way, the notes are not free from typos and grammatical errors!

Contents

1 Universal Approximation Theorem

The following paper answers to the question :

Why do neural networks work at all?

Briefly, the answer is because neural networks with at least one hidden layer and an appropriate activation function can approximate any continuous function arbitrary well. This statement is called the *universal approximation theorem* or the *Cybenko's theorem*. There are many variants of this theorem. We start with the original paper of G.Cybenko.

1.1 Cybenko's theorem.

The original version of the universal approximation theorem appears in [2]. This paper proves that the set of neural networks with a sigmoidal function as an activation function is dense in the set of continuous functions on the n-dimensional unit cube $I_n = [0, 1]^n$. The author shows the statement

in the following scheme:

$$\begin{array}{c} \text{Define a } \textbf{discriminatory} \text{ function.} \\ \downarrow \\ \text{Prove the statement for } \textbf{discriminatory} \text{ functions.} \\ \downarrow \\ \text{Prove that 'good' } \textit{sigmoidal} \text{ functions are } \textbf{discriminatory}. \\ \downarrow \\ \text{Neural networks with a 'good' sigmoidal function as activation function work fine!} \end{array} \tag{1.1}$$

Now let's dive into the paper!

1.1.1 Notation

- I_n = the unit cube $[0,1]^n$.
- $C(I_n)$ = the space of continuous functions on I_n equipped with the supremum norm $\|\cdot\|$.
- $M(I_n)$ = the space of finite, signed, regular Borel measures on I_n.

As we haven seen in the scheme 1.1, we start by defining a **discriminatory** function.

Definition 1.2 ([2])**.** We say that σ is *discriminatory* if for a measure $\mu \in M(I_n)$

$$\int_{I_n} \sigma(y^T x + \theta) d\mu(x) = 0$$

for all $y \in \mathbb{R}^n$ implies that $\mu = 0$.

Remark 1.3. I am not sure if it was the real motivation for the name 'discriminatory'. But, for the moment, it is pretty helpful to think that this property is called **discriminatory** because functions with this property are 'useful' in this paper and others are 'not useful'.

Definition 1.4 ([2])**.** We say that σ is *sigmoidal* if

$$\sigma(t) = \begin{cases} 1 & \text{as } t \to +\infty, \\ 0 & \text{as } t \to -\infty. \end{cases} \tag{1.5}$$

Theorem 1.6 ([2],Theorem 1.)**.** *Let σ be any continuous discriminatory function. Then finite sums of the form*

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \theta_j) \tag{1.7}$$

are dense in $C(I_n)$.

It is worth to read the proof once. The proof is neither that long nor that complicated. So don't worry!

Proof. Let $S \subseteq C(I_n)$ be the set of functions of the form $G(x)$ as in eq. (1.7).

Claim: S is dense in $C(I_n)$.

Assume that it is not the case. Then the closure of S is a proper subset of $C(I_n)$. By the Hahn-Banach theorem, there is a bounded linear functional $L : C(I_n) \to \mathbb{R}$, with the property that $L \neq 0$ but $L(\bar{S}) = L(S) = 0$. (You can find more detailed explanation about Hahn-Banach theorem in A.16)

By the Riesz Representation Theorem (See A.11), this bounded linear functional, L,is of the form

$$L(h) = \int_{I_n} h(x)d\mu(x)$$

for some $\mu \in M(I_n)$, for all $h \in C(I_n)$. Since $L(\bar{S}) = L(S) = 0$, we have that

$$\int_{I_n} \sigma(y^T x + \theta)d\mu(x) = 0$$

for all y and θ. By definition 1.2, it follows that $\mu = 0$, equivalently $L = 0$. This is a contradiction to the assumption that $L \neq 0$. So S is dense in $C(I_n)$. □

Now we show that any continuous sigmoidal function is discriminatory which will lead us to the universal approximation theorem.

Proposition 1.8 ([2], Lemma 1.)**.** *Any bounded, measurable sigmoidal function σ is discriminatory. In particular, any continuous sigmoidal function is discriminatory.*

Proof. Note that for any x, y, θ, ϕ we have

$$\sigma(\lambda(y^T x + \theta) + \phi) \to \begin{cases} 1 & \text{for } y^T x + \theta > 0 \text{ as } \lambda \to \infty, \\ 0 & \text{for } y^T x + \theta < 0 \text{ as } \lambda \to -\infty, \\ \sigma(\phi) & \text{for } y^T x + \theta > 0 \text{ for all } \lambda. \end{cases} \tag{1.9}$$

Thus, the functions $\sigma_\lambda(x) = \sigma(\lambda(y^T x + \theta) + \phi)$ converge pointwise and boundedly to the function

$$\gamma(x) = \begin{cases} 1 & \text{for } y^T x + \theta > 0, \\ 0 & \text{for } y^T x + \theta < 0, \\ \sigma(\phi) & \text{for } y^T x + \theta = 0, \end{cases} \tag{1.10}$$

as $\lambda \to \infty$. For the function $\psi_{y,\theta} = y^T x + \theta$, let $Z_{y,\theta} = \{x | \psi_{y,\theta} = 0\}$, $P_{y,\theta} = \{x | \psi_{y,\theta} > 0\}$. (The names $Z_{y,\theta}$ and $P_{y,\theta}$ stand for '**Z**ero' and '**P**ositive'.) Then by the Lebesgue Bounced Convergence Theorem (See A.5), we have that

$$\begin{aligned} 0 &= \int_{I_n} \sigma_\lambda(x)d\mu(x) \\ &= \int_{I_n} \gamma(x)d\mu(x) \\ &= \sigma(\phi)\mu(Z_{y,\theta}) + \mu(P_{y,\theta}) \end{aligned} \tag{1.11}$$

for all ϕ, θ, y.

Claim : If $\mu(Z_{y,\theta}) = 0$, then $\mu = 0$.

Fix y. For a bounded measurable function h, define the linear functional $F : C(I_n) \to \mathbb{R}$ as

$$F(h) = \int_{I_n} h(y^T x)d\mu(x)$$

and note that F is a bounded functional on $L^\infty(\mathbb{R})$ since μ is a **finite** signed measure. Let $h : \mathbb{R} \to \mathbb{R}$ be a function defined as:

$$h(u) = \begin{cases} 1 & u \geq \theta, \\ 0 & u < \theta. \end{cases} \tag{1.12}$$

We call such a function as an *indicator function* of the interval $[0, \infty)$ Then we have

$$F(h) = \int_{I_n} h(y^T x)d\mu(x) = \mu(Z_{y,-\theta}) + \mu(P_{y,-\theta}) = 0. \tag{1.13}$$

Why is the equation (1.13) true? The reason is because of the equation (1.11) and the fact that σ is continuous so that there is a ϕ that $\sigma(\phi) = 1$.

Similary, $F(h) = 0$, if h is the indicator function of the interval (θ, ∞). Since $F(h+g) = F(h)+F(g)$, $F(h)$ is 0 for any indicator function of any interval and hence for any linear combination of indicator functions, in other words, for any simple functions. Using the fact that simple functions are dense in $L^\infty(\mathbb{R})$, $F = 0$.

In particular, the bounded measurable functions $s_m(u) = \sin(m \cdot u)$ and $c_m(u) = \cos(m \cdot u)$ give

$$F(c + \mathrm{i}s) = \int_{I_n} \cos(m^T x) + \mathrm{i}\sin(m^T x) d\mu(x) = \int_{I_n} \exp(\mathrm{i}m^T x) d\mu(x) = 0 \tag{1.14}$$

for all m. Thus, the fourier transform of μ is 0 which implies that $\mu = 0$. So σ is discriminatory. □

1.1.2 Application to Artificial Neural Networks

The above statements show that neural networks with one internal layer an an arbitrary continuous sigmoidal function can approximate continuous functions arbitrary well providing that no constraints are placed on the number of nodes or the size of the weights. This is the theorem 1.15 below.

Theorem 1.15. *Let σ be any continuous sigmoidal function. Then finite sums of the form*

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \theta_j) \tag{1.16}$$

are dense in $C(I_n)$. In other words, given any $f \in C(I_n)$ and $\varepsilon > 0$, there is a sum $G(x)$ of the form in eq 1.16 so that

$$|G(x) - f(x)| < \varepsilon \quad \text{for all } x \in I_n. \tag{1.17}$$

This theorem implies that neural networks can approximate any classifier arbitrary well.

Theorem 1.18. *Let σ be a continuous sigmoidal function. Let f be the decision function for any finite measurable partition of I_n. For any $\varepsilon \geq 0$, there is a finite sum of the form*

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \theta_j) \tag{1.19}$$

and a set $D \subset I_n$, so that $m(D) \geq 1 - \varepsilon$ and

$$|G(x) - f(x)| < \varepsilon \qquad \text{for } x \in D. \tag{1.20}$$

Proof. Using the Lusin's theorem (See A.7) with the continuous function f, there is a continuous function h and a set D with $\mu(D) \geq 1 - \varepsilon$ so that $h(x) = f(x)$ on D. Now h is continuous and so, by the Theorem 1.15, there is an appropriate G in (1.16) which satisfies $|G(x) - h(x)| < \varepsilon$ for all $x \in I_n$. Then for $x \in D$, we have

$$|G(x) - f(x)| = |G(x) - h(x)| < \varepsilon \tag{1.21}$$

□

Furthermore, the author shows in the paper that artificial neural networks with other activation functions work also fine.

The universal approximation theorem remains also true for

- bounded measurable sigmoidal function σ,
- general sigmoidal function σ.

A Mathematical Preliminaries.

A From Topology.

Definition A.1 (General topological space)**.** Let X be a topological space. A subset $D \subseteq X$ is called *dense* in X if its closure is X.

Definition A.2 (Metric space)**.** Let X be a metric space. A subset $D \subseteq X$ is *dense* if for any $x \in X$ there exists a sequence x_n in D converging to x. In other words, for any $\varepsilon > 0$, there is a $n \in \mathbb{N}$ such that

$$d(x_m, x) < \varepsilon \tag{A.3}$$

for all $m \geq n$.

Example A.4. The set $\mathbb{Q}$ of all rational numbers is dense in $\mathbb{R}$ since for any real number r, there is a sequence of rational numbers that converges to r with respect to the metric

$$d(x, y) = |x - y|.$$

B From measure theory.

Theorem A.5 (Lebesgue Dominated Convergence Theorem)**.** *Let $\{f_n\}$ be a sequence of integrable functions such that*

1. *$f_n \to f$*
2. *there exists a non-negative integrable function g such that $|f_n| \leq g$ for all n.*

Then f is also integrable and

$$\int f_n \xrightarrow[n\to\infty]{} \int f \tag{A.6}$$

Theorem A.7 (Lusin's theorem)**.** *Let X be a locally compact Hausdorff space and $C_c(X)$ be the space of continuous functions on X with compact support. Suppose f is a complex measurable function on X, $\mu(A) < \infty$, $f(x) = 0$ if $x \notin A$, and $\varepsilon > 0$. Then there exists a $g \in C_c(X)$ such that*

$$\mu(\{x : f(x) \neq g(x)\}) < \varepsilon. \tag{A.8}$$

Further more, we may arrange it so that

$$\sup_{x\in X} |g(x)| \leq \sup_{x\in X} |f(x)|. \tag{A.9}$$

Theorem A.10 (Classical Lusin's theorem)**.** *For an interval $[a, b]$, let $f : [a, b] \to \mathbb{C}$ be a measurable function. Then, for every $\varepsilon > 0$, there exists a compact $E \subseteq [a, b]$ such that f restricted to E is continuous and $\mu(E) > b - a - \varepsilon$.*

This theorem states that an almost-everywhere finite measurable function is measurable if and only if it is a continuous function on nearly all its domain.

C From functional analysis.

Let $\mathcal{H}$ be a topological vector space. Throughout this paper, a *linear manifold* is a linear subspace of $\mathcal{H}$ that is not neccesarily closed. A *linear subspace* of $\mathcal{H}$ will always mean a closed linear space.

Theorem A.11 (The Riesz representation theorem)**.** *Let $\mathcal{H}$ be a hilbert space and $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$. If $L : \mathcal{H} \to \mathbb{F}$ is a bounded linear functional, then there is a unique vector h_0 in $\mathcal{H}$ such that $L(h) =< h, h_0 >$ for every $h \in \mathcal{H}$. Moreover, $\|L\| = \|h_0\|$*

Corollary A.12. *If (X, Ω, μ) is a measure space and $F : L^2(\mu) \to \mathbb{F}$ is a bounded linear functional, then there is a unique h_0 in $L^2(\mu)$ such that*

$$F(h) = \int_X h\bar{h_0} d\mu \tag{A.13}$$

for every h in $L^2(\mu)$.

Theorem A.14 ([1],Hahn-Banach)**.** *Let $\mathcal{X}$ be a vector space over $\mathbb{R}$ and let q be a sublinear functional on $\mathcal{X}$. If $\mathcal{M}$ is a linear manifold in $\mathcal{X}$ and $f : \mathcal{M} \to \mathbb{R}$ is a linear functional such that $f(x) \leq q(x)$ for all $x \in \mathcal{M}$, then there is a linear functional $F : \mathcal{X} \to \mathbb{R}$ such that $F|\mathcal{M} = f$ and $F(x) \leq q(x)$ for all x in $\mathcal{X}$.*

Corollary A.15. *[1] If $\mathcal{X}$ is a normed space and $\mathcal{M}$ is a linear manifold in $\mathcal{X}$, then $\mathcal{M}$ is dense in $\mathcal{X}$ if and only if the only bounded linear functional on $\mathcal{X}$ that annihilates $\mathcal{M}$ is the zero functional.*

Remark A.16. Suppose that a linear manifold M is not dense in $\mathcal{X}$. Then the closure N of M is also not dense in $\mathcal{X}$. Then, by Corollary A.15, there is a bounded nonzero linear functional on $\mathcal{X}$ that annihilates N. In other words, there is a bounded linear functional on $\mathcal{X}$, call it L, with the property that $L \neq 0$ but $L(N) = L(M) = 0$.

References

[1] J.B. Conway. *A Course in Functional Analysis*. Graduate Texts in Mathematics. Springer New York, 1994.

[2] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[3] G.B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 2013.

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[5] Friedrich Hirzebruch. *Einführung in die Funktionalanalysis*. B.I. Hochschultaschenbücher, 1971. `https://hirzebruch.mpim-bonn.mpg.de/id/eprint/117/`.