A Theoretical Analysis of (Artificial) Neural Networks

Daniele Kopyshevskiy

June 6, 2025

Contents

Introduction **2**

1 Sigmoidal functions **3**

2 Universal approximation theorem **3**

3 Complexity of NN **7**

Introduction

The empirical success of neural networks in fields ranging from computer vision to natural language processing is well-documented. Underpinning this success is a rich mathematical theory that explains the fundamental capabilities and limitations of these models. These notes provide a concise overview of some of the cornerstone theoretical results concerning the expressive power of neural networks.

The analysis begins with the foundational components of a network: the activation functions. We formally define sigmoidal functions and the crucial property of being "discriminatory." These properties are essential for establishing the broader capabilities of the network architectures they enable.

The centerpiece of the discussion is the **Universal Approximation Theorem**. This theorem addresses the fundamental question: can a neural network with a single hidden layer approximate any continuous function to an arbitrary degree of accuracy? We will explore a proof of this theorem that leverages fundamental results from functional analysis, such as the Hahn-Banach and Riesz Representation theorems, to demonstrate that the space of functions representable by a shallow network is dense in the space of continuous functions on a compact domain. We will show how this property holds for both traditional sigmoidal activations and the widely-used ReLU function.

Finally, building upon this existence result, we turn to the question of **complexity and efficiency**. While the Universal Approximation Theorem guarantees that a function *can* be approximated, it does not specify the size of the network required. We will present key theorems that quantify the number of neurons (the network's complexity) needed to achieve a given approximation accuracy, ϵ. This analysis contrasts the performance of shallow and deep networks, revealing the theoretical advantage of deep architectures in mitigating the "curse of dimensionality" for certain classes of functions.

1 Sigmoidal functions

Definition 1.1. A function $\sigma : \mathbb{R} \to [0, 1]$ is called sigmoidal if

$$\lim_{x \to -\infty} \sigma(x) = 0, \qquad \lim_{x \to +\infty} \sigma(x) = 1.$$

Definition 1.2. Let n be a natural number and $I_n = [0, 1]^n$. We say that an activation function $f : \mathbb{R} \to \mathbb{R}$ is n-discriminatory if the only signed Borel measure μ such that

$$\int_{I_n} f(y \cdot x + \theta) d\mu(x) = 0, \quad \forall y \in \mathbb{R}^n, \theta \in \mathbb{R},$$

is the zero measure.

Definition 1.3. We say an activation function $f : \mathbb{R} \to \mathbb{R}$ is discriminatory if it is n-discriminatory for any n.

Remark 1.1. A discriminatory function σ is volumetrically non-destructive when it acts on linear transformations of input.

2 Universal approximation theorem

Let x be the input variable, z the target and denote the target function by $z = f(x)$, with f in a certain function space S.

Some definitions:

- $I_n = [0, 1]^n$;
- A subspace U of X is *dense* in X with respect to a norm $\| \cdot \|$ if for any element $x \in X$ there are elements $u \in U$ as close as possible to x. Alternatively:
  1. $\forall x \in X$ there is a sequence u_n in U s.t. $u_n \to x$, as $n \to \infty$;
  2. $\forall x \in X, \forall \epsilon > 0$, there is $u \in U$ s.t. $\|u - x\| < \epsilon$
- The fact that the subspace U is *not dense* in X can be described as:
  1. there are elements $x_0 \in X$ s.t. no elements $u \in U$ are close enough to x_0;
  2. there is a $\delta > 0$ s.t. $\forall u \in U$ we have $\|u - x_0\| \geq \delta$.
- We say that the neural network is a *universal approximator* for the space (S, d) if the space of outcomes U is d-dense in S *i.e.*

  $$\forall f \in S, \quad \forall \epsilon > 0, \quad \exists g \in U : d(f - g) < \epsilon$$

  In practice it means that for any function $f \in S$, there are functions in U situated in any proximity of f.
- let K denote a compact set in $\mathbb{R}^n$ and denote by $C(K)$ the set of real-valued continuous function on K;
- $M(I_n)$ is the space of finite signed regular Borel measure on I_n. (*Note:* regular means that while different measures may define different sizes for a single set, they all ideally convey some idea of how much space that set takes up relative to the larger space in which it resides).

Theorem 2.1 (Representation of Linear Bounded Functional)**.** *Let F be a bounded linear functional on $C(K)$. Then there exists a unique finite signed Borel measure μ on K such that*

$$F(f) = \int_K f(x)d\mu(x), \quad \forall f \in C(K).$$

Moreover $\|F\| = |\mu|(K)$.

Theorem 2.2 (Hahn-Banach)**.** *Let X be a linear real vector space, X_0 a linear subspace, p a linear convex functional on X, and $f : X_0 \to \mathbb{R}$ a linear functional s.t. $f(x) \leq p(x)$ for all $x \in X_0$. Then there is a linear functional $F : X \to \mathbb{R}$ s.t.:*

1. *$F_{X_0} = f$ (the restriction of F to X_0 is f)*
2. *$F(x) \leq p(x)$ for all $x \in X$.*

Remark 2.1. The Hahn-Banach theorem tells us that we can always extend a linear functional defined on a subspace to obtain a linear functional on the whole space that behaves in the same way. This is useful because it allows us to study the behavior of linear functionals on a larger space, which can provide more information about the structure of the original subspace.

From the Hahn-Banach theorem we have the following two lemmas.

Lemma 2.1. *Let U be a linear subspace of a normed linear space X and consider $x_0 \in X$ such that*

$$\operatorname{dist}(x_0, U) \geq \delta,$$

for some $\delta > 0$, i.e.,

$$\|x_0 - u\| \geq \delta, \forall u \in U.$$

Then there is a bounded linear functional L on X such that:

(i) $\|L\| \leq 1$

(ii) $L(u) = 0, \quad \forall u \in U$, *i.e.,* $L_{|U} = 0$

(iii) $L(x_0) = \delta$.

Lemma 2.2 (Reformulation of Lemma 1)**.** *Let U be a linear, nondense subspace of a normed linear space X. Then there is a bounded linear functional L on X such that $L \neq 0$ on X and $L_U = 0$.*

Lemma 2.3. *Let U be a linear, non-dense subspace of $C(I_n)$. Then there is a measure $\mu \in M(I_n)$ such that*

$$\int_{I_n} h d\mu = 0, \quad \forall h \in U.$$

Proof. Considering $X = C(I_n)$ in Lemma 2, there is a bounded linear functional $L : C(I_n) \to \mathbb{R}$ such that $L \neq 0$ on $C(I_n)$ and $L_{|U} = 0$. Applying the representation theorem of linear bounded functionals on $C(I_n)$ there is a measure $\mu \in M(I_n)$ such that

$$L(f) = \int_{I_n} f d\mu, \quad \forall f \in C(I_n).$$

In particular for any $h \in U$ we have

$$L(h) = \int_{I_n} h d\mu = 0,$$

which is the desired result. □

Remark 2.2. Note that $L \neq 0$ implies $\mu \neq 0$.

Proposition 2.1. *Let σ be any continuous discriminatory function. Then the finite sums of the form*

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma\left(w_j^T x + \theta_j\right), \quad w_j \in \mathbb{R}^n, \alpha_j, \theta_j \in \mathbb{R}$$

are dense in $C(I_n)$.

Proof. Since σ is continuous, it follows that

$$U = \left\{ G; G(x) = \sum_{j=1}^{N} \alpha_j \sigma\left(w_j^T x + \theta_j\right) \right\}.$$

is a linear subspace of $C(I_n)$. We continue the proof adopting the contradiction method.

Assume that U is not dense in $C(I_n)$ *i.e.* we assume that the closure of U is not all $C(I_n)$. Then the closure of U (call it R) is a closed proper subspace of $C(I_n)$.

By the Hahn-Banach Theorem there is a bounded linear functional on $C(I_n)$ (call it L) with the property that $L \neq 0$ but $L(U) = L(R) = 0$.

By the Representation Theorem L is of the form

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

for some $\mu \in M(I_n)$, for all $h \in C(I_n)$.

In particular since $\sigma(w^T x + \theta)$ is in U for all w and θ, we must have

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

However we have assumed that σ was discriminatory so this implies $\mu = 0$ which contradicts our assumption; hence S must be dense in $C(I_n)$. □

Definition 2.1. Let

- $P_{w,\theta} = \left\{x; w^T x + \theta = 0\right\}$ the hyperplane with normal vector w and intercept θ;
- $H^+_{w,\theta} = H_{w,\theta} = \left\{x; w^T x + \theta > 0\right\}$ the positive half-space;
- $H^-_{w,\theta} = \left\{x; w^T x + \theta < 0\right\}$ the negative half-space.

Lemma 2.4. *Let $\mu \in M(I_n)$. If μ vanishes on all hyperplanes and open half-spaces in $\mathbb{R}^n$ then μ is zero. More precisely if*

$$\mu(P_{w,\theta}) = 0, \quad \mu(H_{w,\theta}) = 0, \quad \forall w \in \mathbb{R}^n, \theta \in \mathbb{R},$$

then $\mu = 0$.

Proposition 2.2. *Any continuous sigmoidal function is discriminatory for all measures $\mu \in M(I_n)$.*

Proof. Let $\mu \in M(I_n)$ be a fixed measure. Choose a continuous sigmoidal function that satisfies

$$\int_{I_n} \sigma(w^T x + \theta) d\mu(x) = 0, \quad \forall w \in \mathbb{R}^n, \theta \in \mathbb{R} \tag{1}$$

We need to show that $\mu = 0$. First, construct the continuous function

$$\sigma_\lambda(x) = \sigma\left(\lambda\left(w^T x + \theta\right) + \phi\right)$$

for given w, θ and ϕ, and use the definition of a sigmoidal to note that

$$\lim_{\lambda\to\infty} \sigma_\lambda(x) = \begin{cases} 1, & \text{if } w^T x + \theta > 0 \\ 0, & \text{if } w^T x + \theta < 0 \\ \sigma(\phi), & \text{if } w^T x + \theta = 0 \end{cases}$$

Define the bounded function

$$\gamma(x) = \begin{cases} 1, & \text{if } x \in H^+_{w,\theta} \\ 0, & \text{if } x \in H^-_{w,\theta} \\ \sigma(\phi), & \text{if } x \in P_{w,\theta} \end{cases}$$

and notice that $\sigma_\lambda(x) \to \gamma(x)$ pointwise on $\mathbb{R}$, as $\lambda \to \infty$. The Bounded Convergence Theorem allows switching the limit with the integral, obtaining

$$\begin{aligned} \lim_{\lambda\to\infty} \int_{I_n} \sigma_\lambda(x) d\mu(x) &= \int_{I_n} \gamma(x) d\mu(x) \\ &= \int_{H^+_{w,\theta}} \gamma(x) d\mu(x) + \int_{H^-_{w,\theta}} \gamma(x) d\mu(x) + \int_{P_{w,\theta}} \gamma(x) d\mu(x) \\ &= \mu\left(H^+_{w,\theta}\right) + \sigma(\phi)\mu\left(P_{w,\theta}\right). \end{aligned}$$

Equation (1) implies that $\int_{I_n} \sigma_\lambda(x) d\mu(x) = 0$, and hence the limit in previous left term vanishes. Consequently, the right term must also vanish, fact that can be written as

$$\mu\left(H^+_{w,\theta}\right) + \sigma(\phi)\mu\left(P_{w,\theta}\right) = 0.$$

Since this relation holds for any value of ϕ, taking $\phi \to +\infty$ and using the properties of σ, yields

$$\mu\left(H^+_{w,\theta}\right) + \mu\left(P_{w,\theta}\right) = 0.$$

Similarly, taking $\phi \to -\infty$, implies

$$\mu\left(H^+_{w,\theta}\right) = 0, \quad \forall w \in \mathbb{R}^n, \theta \in \mathbb{R}. \tag{2}$$

Note that, as a consequence of the last two relations, we also have $\mu(P_{w,\theta}) = 0$. Since $H^+_{w,\theta} = H^-_{-w,-\theta}$, relation (2) states that the measure μ vanishes on all half-spaces of $\mathbb{R}^n$. Lemma 4 states that a measure with such properties is necessarily the zero measure, $\mu = 0$. Therefore, σ is discriminatory. □

Proposition 2.3. *The ReLU function is 1-discriminatory.*

Proof. Let μ be a signed Borel measure, and assume the following holds for all $y \in \mathbb{R}$ and $\theta \in \mathbb{R}$:

$$\int \mathrm{ReLU}(yx + \theta) d\mu(x) = 0$$

We want to show that $\mu = 0$. For that, we will construct a sigmoid bounded, continuous (and therefore Borel measurable) function from subtracting two ReLU functions with different parameters. In particular, consider the function

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \in [0, 1] \\ 1 & \text{if } x > 1 \end{cases}$$

Then any function of the form $g(x) = f(yx + \theta)$ with $y \neq 0$ can be described as

$$g(x) = \text{ReLU}\left(yx + \theta_1\right) - \text{ReLU}\left(yx + \theta_2\right)$$

by setting $\theta_1 = -\theta/y$ and $\theta_2 = (1-\theta)/y$. If $y = 0$, then instead set

$$g(x) = f(\theta) = \begin{cases} \text{ReLU}(f(\theta)) & \text{if } f(\theta) \geq 0 \\ -\text{ReLU}(-f(\theta)) & \text{if } f(\theta) \leq 0 \end{cases}$$

Which means that for any $y \in \mathbb{R}, \theta \in \mathbb{R}$

$$\begin{aligned} \int f(yx+\theta)d\mu(x) &= \int \left(\text{ReLU}\left(yx + \theta_1\right) - \text{ReLU}\left(yx + \theta_2\right)\right) d\mu(x) \\ &= \int \text{ReLU}\left(yx + \theta_1\right) d\mu(x) - \int \text{ReLU}\left(yx + \theta_2\right) d\mu(x) \\ &= 0 - 0 = 0. \end{aligned}$$

By the previous lemma, f is discriminatory, and therefore, $\mu = 0$. □

Definition 2.2. For $f : \mathbb{R} \to \mathbb{R}$ an activation function we define:

$$\Sigma_n(f) = \text{span}\left\{f(y \cdot x + \theta) | y \in \mathbb{R}^n, \theta \in \mathbb{R}\right\}.$$

Proposition 2.4. *If $\Sigma_1(f)$ is dense in $C([0,1])$ then $\Sigma_n(f)$ is dense in $C([0,1]^n)$.*

3 Complexity of NN

- W_m^n: class of n-variable functions with partial derivatives up to m-th order
- $W_m^{n,2} \subset W_m^n$: compositional subclass following binary tree structure

Theorem 3.1. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be infinitely differentiable, and not a polynomial. For $f \in W_m^n$ the complexity of shallow networks that provide accuracy at least ϵ is $N = O\left(\epsilon^{-n/m}\right)$ and is the best possible.*

Theorem 3.2. *For $f \in W_m^{n,2}$ consider a deep network with the same compositional architecture and with an activation function $\sigma : \mathbb{R} \to \mathbb{R}$ which is infinitely differentiable, and not a polynomial. The complexity of the network to provide approximation with accuracy at least ϵ is*

$$N = O\left((n-1)\epsilon^{-2/m}\right). \tag{3}$$

Theorem 3.3. *Let f be a L-Lipshitz continuous function of n variables. Then, the complexity of a network which is a linear combination of ReLU providing an approximation with accuracy at least ϵ is*

$$N_s = O\left(\left(\frac{\epsilon}{L}\right)^{-n}\right)$$

whereas that of a deep compositional architecture is

$$N_d = O\left((n-1)\left(\frac{\epsilon}{L}\right)^{-2}\right).$$