Universal Approximation Theorem

Aaron Avram

June 9 2025

Introduction

In this file I will outline a proof of the Universal Approximation theorem inspired by the one published by Cybenko et al. As a disclaimer, I understand the proof of the theorem, however some of the theorems it leverages are ones that are slightly above the scope of my mathematical knowledge. I specifically refer to the geometric Hahn-Banach Theorem and the Riesz-Markov-Kakutani representation Theorem. My familiarity with functional analysis and measure theory is rudimentary, but I plan on spending more time with this material soon and may revisit this proof at that point.

For now let us work through the proof. This is my own formulation of the classic proof structure.

Construction

The classic proof of the Universal Approximation Theorem by Cybenko et al. is done for a neural network with one hidden layer and scalar-valued output. This can be extended to a proof of the more general case where the network has an arbitrary number of hidden layers and vector-valued output. However, I will not comment on the latter case and will focus on the proof of the first case which is the backbone of the more general proof as well.

Now to define our setup more clearly: we consider a classic neural network structure with three layers: the input, hidden and output layer. We can represent the first two layers with vectors $x^{(0)}, x^{(1)}$ then the output layer as a scalar y. We construct each $x_i^{(1)} = \sigma\Big(\sum_{j=1}^{n} x_j^{(0)} w_{ji}^{(1)} + b_i^{(1)}\Big)$. This can be seen as taking a weighted sum of the previous layer, adding the bias then applying a non-linearity to squash the value. More succinctly $x_i^{(1)} = \sigma((x^{(0)})^T w_i^{(1)} + b_i^{(1)})$. Thus it can be observed that each hidden layer neuron is given by a function $g_i : \mathbb{R}^n \to \mathbb{R}$ applied to $x^{(0)}$. Our output layer is merely a linear combination of the hidden layer values $y = \sum_{i=1}^{m} w_i g_i(x^{(0)})$. Then we can write our entire network as a function $\psi(x) : \mathbb{R}^n \to \mathbb{R}$.

The last note before we proceed, is about σ. This is called an activation function and for the proof we want to introduce two constraints on this function. First we will restrict our input space to I_n which is the closed unit n-dimensional hypercube. Let $M(I_n)$ be the set of finite signed regular Borel measures on I_n then:

Definition A function σ is discriminatory if for a measure $\mu \in M(I_n)$, $\int_{I_n} \sigma(w^T x + b) d\mu(x) = 0$, $\forall w \in \mathbb{R}^n$ and $\forall b \in \mathbb{R}$. implies that μ is the zero measure.

We require that σ is discriminatory and while I do not understand the exact technicalities of what this means, intuitively I understand it as the ability to distinguish between different measures (I.e. patterns in the input space) as if two measures μ_1, μ_2 produce the same value for all such integrals then applying the definition to $\mu_1 - \mu_2$ implies that they are in fact the same measure.

We additionally require that σ approaches 0 as it goes to $-\infty$ and 1 as it goes to ∞.

Proof

We specifically want to show that we can approximate any continuous function from I_n to $\mathbb{R}$. Denote the set of such functions as $C(I_n)$. Observe that the function $g(x) = \sigma(x^T w + b)$ for some w and b is in this set. additionally the set of all neural networks, which we will denote $\mathcal{N}$ is the set of linear combinations of functions of the same form as g which is a subspace of $C(I_n)$. Thus, to show that any function $f \in C(I_n)$ can be approximated arbitrarily well by a function in $\mathcal{N}$ we seek $\Psi \in \mathcal{N}$ such that $||\psi - f||_\infty < \epsilon$ where $||.||\infty$ is the supremum norm. This equivalent to showing that $\mathcal{N}$ is dense in $C(I_n)$. We will prove this by contradiction.

Proof. Suppose that $\mathcal{N}$ is not dense in $C(I_n)$ then $\overline{\mathcal{N}} \subsetneq C(I_n)$. Thus, by the Banach-Hahn theorem there exists a functional ϕ on $C(I_n)$ such that ϕ is not zero but $\phi(\overline{\mathcal{N}}) = 0$. Then by the Riesz-Kakutani-Markov representation theorem we can represent this functional with an integral, for some $\mu \in M(I_n)$ by $\phi(h) = \int_{I_n} h(x) d\mu(x)$. However this is true in particular for functions in $\mathcal{N}$. Thus for all w and b and $g := \sigma(x^T w + b)$, $0 = \phi(g) = \int_{I_n} g(x) d\mu(x)$. But, this implies that μ is the zero measure as σ is discriminatory and this in turn implies that ϕ is the zero functional, which is a contradiction. Thus, $\mathcal{N}$ is dense in $C(I_n)$ as required. □