Approximation of BV functions using neural networks

Benny Avelin
Joint work with Vesa Julin (Jyväskylä)

2021

Abstract

In this talk, I will focus on a recent result together with Vesa Julin, concerning the approximation of functions of Bounded Variation (BV) using special neural networks on the unit circle. I will present the motivation for studying these special networks, their properties, and hopefully some proofs. Specifically the results we will cover: the closure of the class of neural networks in L^2, a uniform approximation result, and a localization result.

Origin of the problem

A real valued single hidden layer neural network is defined as

$$f_{W,a}(x) = \sum_{i=1}^{m} a_i \sigma(w_i \cdot x + b_i) \; : \; \Omega \subset \mathbb{R}^n \to \mathbb{R}$$

usually $W \in \mathbb{R}^{m(n+2)}$ is $(a_1, \ldots, a_m, w_1, b_1, w_2, b_2, \ldots)$. Above σ is called an activation function.

1. $\sigma(x) = \frac{1}{1+e^{-x}}$, sigmoid
2. $\sigma(x) = \tanh(x)$
3. $\sigma(x) = \max(0, x)$, ReLU.

Origin of the problem

Neural networks can be used for many things, but often they are used in the context of (least squares) regression

$$\inf_{W,a} \|f_{W,a} - y\|^2_{L^2(\mu)}$$

where μ is the empirical measure and $y(x)$ is the observation at x.

A network on the sphere

Consider

$$f_{W,a}(x) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} a_i \sigma(w_i \cdot x),$$

where the vectors $w_i \in \mathbb{R}^n$, $W = (w_1, \dots, w_m) \in \mathbb{R}^{mn}$, denote the weights, and the coefficients $a_i \in \{-1, 1\}$, $a = (a_1, \dots, a_m)$, are given and the activation function is $\sigma(t) = \max\{t, 0\}$.

Exponential convergence in the overparametrized regime

Let μ be the empirical measure for N datapoints (x_i, y_i), where x_i is on the unit sphere, and $|y_i| \leq 1$.

Theorem (Du, Zhai, Poczós, Singh: 2019)

If the number of hidden nodes $m \gtrsim \frac{N^6}{\lambda_0^4}$ and we intialize W, a randomly, then with high probability we have

$$\|f_{W(t),a} - y\|^2_{L^2(\mu)} \leq e^{-\lambda_0 t} \|f_{W(0),a} - y\|^2_{L^2(\mu)}$$

Exponential convergence in the overparametrized regime

Let μ be the empirical measure for N datapoints (x_i, y_i), where x_i is on the unit sphere, and $|y_i| \leq 1$.

Theorem (Du, Zhai, Poczós, Singh: 2019)

If the number of hidden nodes $m \gtrsim \frac{N^6}{\lambda_0^4}$ *and we intialize* W, a *randomly, then with high probability we have*

$$\|f_{W(t),a} - y\|^2_{L^2(\mu)} \leq e^{-\lambda_0 t}\|f_{W(0),a} - y\|^2_{L^2(\mu)}$$

Here λ_0 is the smallest eigenvalue of the following Gramian matrix

$$\mathbb{E}_{w\sim N(0,1)}\left[(x_i \mathbb{I}\{x_i \cdot w \geq 0\}) \cdot (x_j \mathbb{I}\{x_j \cdot w \geq 0\})\right]$$

What happens in the underparametrized regime?

Recall that the network is (in 2D)

$$f_{W,a}(x) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} a_i \sigma(w_i \cdot x), \quad x \in S^1$$

where the vectors $w_i \in \mathbb{R}^2$, $W = (w_1, \dots, w_m) \in \mathbb{R}^{2m}$, denote the weights, and the coefficients $a_i \in \{-1, 1\}$, $a = (a_1, \dots, a_m)$, are given and the activation function is $\sigma(t) = \max\{t, 0\}$.

Given the vector a we denote the class of functions above as $\mathcal{H}_{m,a}$.

What happens in the underparametrized regime?

These are the questions that we would like to answer in a quantitative way:

1. Do we still have exponential convergence?
2. What can we say about the minimum value of

$$\inf_{W} \|f_{W,a} - y\|^2$$

The behavior changes

Let us take the following example

$$y(x) = y(x_1, x_2) = \mathbb{I}_{\{x_2 \geq 0\}} x_1$$

and consider approximating the above using $\mathcal{H}_{2,a}$, $a = (1, -1)$. We will see in a moment that

$$\inf_{W \in \mathbb{R}^4} \Phi(W) := \inf_{W \in \mathbb{R}^4} \| f_{W,a} - y \|_{L^2(S^1)} = 0,$$

but $\Phi(W) > 0$ for all $W \in \mathbb{R}^4$. Furthermore

$$\frac{d}{dt} W_t = -\nabla_W \Phi(W)$$

satisfies $\lim_{t \to \infty} \|W_t\| = \infty$ for certain W_0.

The behavior changes

$$y(x) = y(x_1, x_2) = \mathbb{I}_{\{x_2 \geq 0\}} x_1$$

Take the network

$$f_{W,a}(x) = \sigma\left(\frac{1}{h}x_2 + x_1\right) - \sigma\left(\frac{1}{h}x_2\right)$$

The behavior changes

So,

1. the problem is not coercive
2. does not have a global minimum
3. and the gradient descent may diverge.

A way to get around this would be to consider a penalized form of the minimization problem, to keep $|W|$ bounded.

The goal

1. What can we say quantitatively about the value of

$$\inf_{W} \|f_{W,a} - y\|_{L^2}^2 \,?$$

2. And how far away from that is

$$\inf_{W \in B_R} \|f_{W,a} - y\|_{L^2}^2 \,?$$

   In other words, how much do we pay in terms of the minimum value, in order to constrain the minimization problem?

When we say quantitative, we mean estimates with explicit constants.

Closure in L^2

Our example shows that in general there is no global minima for

$$\inf_W \|f_W - y\|_2^2$$

so we need to identify the closure of $\mathcal{H}_{m,a}$ in L^2.

Theorem

A function $g : S^1 \to \mathbb{R}$ belongs to the space $\overline{\mathcal{H}}_{m,a}$ if and only if it is of the form

$$g(x) = \sum_{i \in J} \mathbb{I}\{\hat{w}_i \cdot x \geq 0\}(v_i \cdot x) + \sum_{i \in K} a_i \sigma(w_i \cdot x)$$

where $\hat{w}_i$ are unit vectors, the set of indices J, K are disjoint and $|J| \leq \underline{m}$.

$$\underline{m} = \min\{|i : a_i = -1|, |i : a_i = 1|\}$$

The properties of the function class

Simple observation

$$\sigma(t) = \max\{t, 0\} = \frac{|t|}{2} + \frac{t}{2} = \text{symmetric} + \text{linear}.$$

since $w_i \cdot x$ is linear, we know that every function $f_W \in \mathcal{H}_{m,a}$ is

$$f_W = \text{antipodally symmetric} + \text{linear}.$$

Symmetry

Lemma

For a function on S^1 of the form

$$g(x) = \sum_{i \in J} \mathbb{I}\{\hat{w}_i \cdot x \geq 0\} v_i \cdot x$$

if we decompose it into the antipodally symmetric and antisymmetric parts we get

$$g_s(x) = \frac{1}{2} \sum_{i \in J} sgn(\hat{w}_i \cdot x)(v_i \cdot x)$$

$$g_a(x) = \frac{1}{2} \left(\sum_{i \in J} v_i \right) \cdot x$$

Uniform approximation theorem

Written in polar coordinates, the function $g : S^1 \to \mathbb{R}$

$$g(\theta) = \mathbb{I}_{(-\theta_0,\theta_0)} \cos(\theta) - \mathbb{I}_{(-\theta_0+\pi,\theta_0+\pi)} \cos(\theta) \in \overline{\mathcal{H}}_{2,(-1,1)}$$

Theorem (Uniform approximation)

Assume that $y \in BV(S^1)$ is symmetric + linear, then

$$\inf_{f_W \in \mathcal{H}_{m,a}} \|f_W - y\|_2^2 \leq \frac{62\|y\|_{BV}^2}{\underline{m}}.$$

Main result, the localization theorem

Theorem

Assume that $y \in BV(S^1)$ is such that $\|y\|_{L^2(S^1)} \leq 1$. Then for all $R \geq R_0$ the following holds

$$\min_{\substack{f_W \in \mathcal{H}_{m,a} \\ |W| \leq C(m)R}} \|f_W - y\|^2_{L^2(S^1)}$$

$$\leq \inf_{f_W \in \mathcal{H}_{m,a}} \|f_W - y\|^2_{L^2(S^1)} + 5 \cdot 10^4 (\|y\|^2_{BV} + 1) \frac{1}{R^{1/9}},$$

where $C(m) = \sqrt{m/\underline{m}}$ and

$$R_0 = \max\{(10\|y\|_{BV})^6, 4 \cdot 10^7\}.$$

Idea of proof

1. We prove it assuming first that $y \in C^1(S^1)$.
2. The closure allows us to find the minimizer

$$g(x) = \sum_{i \in J} \mathbb{I}\{\hat{w}_i \cdot x \geq 0\}(v_i \cdot x) + \sum_{i \in K} a_i \sigma(w_i \cdot x)$$

3. The symmetric part of the minimizer satisfies an Euler-Lagrange equation. The Euler-Lagrange equation can be used to prove that the minimizer inherits some regularity from y.
4. We then turn these regularity estimates into bounds of the vectors v_i.
5. We use these bounds to construct a local approximation $|W| \leq C(m)R$.

Thank you