The Effect of Higher Order Activation Functions on Infinitely wide Neural Networks

Tjeerd Jan Heeringa, s1497324

2022-04-18

Abstract

Machine learning using neural networks is a very powerful tool used for solving high dimensional and nonlinear problems. Neural networks can approximate almost any function to arbitrary precision, and seem not to suffer from the curse of dimensionality. A key goal in Applied Mathematics and Computer Science is to understand which neural networks can approximate which functions well, and to figure out why neural networks do not suffer from the curse of dimensionality. A major step towards that goal is to understand the continuous limit of neural networks. In recent research several function spaces have been suggested as candidates for this limit. However, the mathematical relationship between these spaces is not fully understood. In the thesis we further the understanding of these candidates, and investigate which of the function spaces continuously embeds into which. We show that Barron space with ReLU as activation function is the largest of these spaces, and derive a novel description of the remainder of a Taylor series in terms of a shallow neural network with the higher order ReLU as activation function. We demonstrate that this shallow neural network does not suffer from the curse of dimensionality. We conclude with an analysis of the continuous limit of a deep neural network using control techniques.

Contents

1 Introduction

In many real world scenarios we have access to data, and we are interested in the structure behind this data. We wish to find a function that represents the data well enough, whilst simultaneously predicts any new data we might acquire accurately. What function does this best depends on the scenario. Typically we will consider a set of candidate functions, and try to find the best function among those. This set of candidate functions needs to be chosen carefully. For example trying to represent a sinusoidal function using a linear function is generally not going to work well.

To determine which set of candidate functions is a good choice, we study the properties of these sets. This is done by studying error bounds. One of these errors is the approximation error, where a bound is sought for how well a candidate function f can be approximated by an approximation of it, f_m, that is only allowed to use a finite number of parameters, m. When norms can be used, the error bounds are often given in the form of an a-priori or a-postiori error bound. For the approximation error between f and f_m these are given by

$$\|f - f_m\|_1 \leqslant C_1 m^{-\alpha} \|f\|_2, \qquad \text{A-priori} \tag{1}$$

$$\|f - f_m\|_3 \leqslant C_2 m^{-\beta} \|f_m\|_4. \qquad \text{A-postiori} \tag{2}$$

Here $\|\cdot\|_i$ are norms that depend on which normed vector spaces f and f_m are from, and C_1, C_2, α, β are constants that depend on the problem. The higher m needs to be to let f_m properly approximate f, the higher the computational cost. Hence, we want C_1 and C_2 to be small and more importantly α and β to be big. One problem dependent parameter that influences the constants C_1, C_2, α, β is the dimension d; α and β often become small when d increases. A set of candidate functions is said to suffer from the *curse of dimensionality* when this happens. Such a set is likely of little use when the problem studied is a high dimensional one, like image recognition. To illustrate this let $f \in \mathcal{H}^\alpha$ be a function from a Sobolev space and f_m an approximation of f consisting of m piece-wise polynomials. The a-postiori approximation error bound is given by

$$\|f - f_m\|_{L^2} \leqslant C m^{-\alpha/d} \|f\|_{H^\alpha} \tag{3}$$

for some C. To achieve $m^{-\alpha/d} = 0.1$, we need $m = 10^{d/\alpha}$. Hence, we need exponentially more parameters to get to the same level of accuracy for higher d [E et al., 2020]. This is alright for low dimensional problems, but if the dimension is high then this becomes prohibitively expensive.

Empirical evidence has shown that neural networks do not suffer from the curse of dimensionality. Hence, we should be able to construct a function space that contains all neural networks, and provide an a-priori or a-postiori bound that shows the curse of dimensionality is indeed not present. Neural networks come in many forms. This means that finding a function space for all of them is a non-trivial matter. Hence, in this work we will mainly talk about the simplest type of neural networks. These are 2 layer neural networks with a single output, called shallow neural networks. When a shallow neural network has m neurons in the hidden layer, it can be represented using

$$f_m(x) = c + \sum_{i=1}^{m} a_i \phi(\langle x | w_i \rangle + b_i), \tag{4}$$

where a_i and c are the weights and bias associated to the output node, w_i and b_i are the weights and biases for the hidden layer, and ϕ the activation function for the hidden layer. The activation

function is applied pointwise. Typically the constant c is omitted, since it can be fitted using a single extra neuron in the hidden layer for all commonly used activation functions. We will also do this in this work. All shallow neural networks with m elements can be grouped into a set F_m. This is not a vector space, since $F_m + F_m \subseteq F_{2m}$. This shows that we should look to a bigger set of function, and consider shallow neural networks approximations thereof.

1.1 Related work

One of the function spaces that has been associated to shallow neural networks is the class of the functions that satisfy

$$\|f\|_{\mathscr{F}^{s,1}} = \int_{\mathbb{R}^d} \|\xi\|^s \left|\hat{f}(\xi)\right| d\xi < \infty, \tag{5}$$

where $\hat{f}$ is the Fourier transform of f [Barron, 1993]. Using a separation argument Barron showed that any function $f \in \mathscr{F}^{1,1}$ could be approximated with shallow neural networks f_m using the step function as activation function with the bound

$$\|f - f_m\|_{L^2(\mathcal{X},\pi)} \leqslant C \frac{\|f\|_{\mathscr{F}^{1,1}}}{\sqrt{m}} \tag{6}$$

for some $C > 0$, compact set $\mathcal{X} \subseteq \mathbb{R}^d$ and probability measure π. This shows that for $\mathscr{F}^{s,1}$ the curse of dimensionality was avoided. Furthermore, by allowing the weights to become arbitrarily large he showed that a similar result also holds for any sigmoidal activation function. Later he proved a similar result with ReLU as activation function for functions $f \in \mathscr{F}^{2,1}$, and up to a linear correction with a squared ReLU as activation function for functions $f \in \mathscr{F}^{3,1}$[Klusowski and Barron, 2018].

The bound achieved by Barron is a Monte Carlo bound. This inspired E. et al. to slightly modify the shallow Neural Network by adding a prefactor $\frac{1}{m}$, so that the shallow neural network, now given by

$$f_m(x) = \frac{1}{m} \sum_{i=1}^{m} a_i \phi(\langle x|w_i\rangle + b_i), \tag{7}$$

can be seen as the empirical estimate of the expectation

$$f(x) = \int_\Omega a\phi(\langle x|w\rangle + b) d\pi(a,w,b) = \mathbb{E}_{(a,w,b)\sim\pi}[a\phi(\langle x|w\rangle + b)]. \tag{8}$$

The space constructed using these expectations is called Barron space. It has been shown that this space produces the same Monte Carlo bound as Barron[E et al., 2021; theorem 4], that it contains all functions from $\mathscr{F}^{2,1}$ when ϕ is ReLU [E et al., 2021; proposition 2], and that every shallow neural network with an activation function satisfying

$$\int_{\mathbb{R}} \left|\partial^2 \phi(x)\right| (1 + |x|) dx < \infty \tag{9}$$

can be approximated by a shallow neural network with ReLU as activation function, which is an element of Barron space[Li et al., 2020].

1.2 Our contribution

In this thesis we further the understanding of this Barron space. Our main focus is on how Barron space changes when we change the activation function, and on the relation between Barron spaces and other spaces. Additionally, we discuss how machine learning can be applied in control, and study a generalisation of Barron spaces using control techniques.

In our study of the effect of the activation function on Barron spaces we had a particular interest in the higher order ReLU

$$\sigma_s(x) = \begin{cases} x^s & x \geqslant 0 \\ 0 & x < 0 \end{cases}$$

with $s \in \mathbb{N}$. Two novel results are derived that use the identity

$$\int_0^z (z-u)^s f(u)du = \int_0^c \sigma_s(z-u)f(u) + (-1)^{s-1}\sigma_s(-z-u)f(-u)du \tag{10}$$

where $z \in [-c, c]$ and f integrable on $[-c, c]$. The key idea of this identity is that it removes the dependence on z from the limits of integration, so that the right hand side can be interpreted as a Barron function with the higher order ReLU as activation function.

The first novel result is an interpretation of the remainder of Taylor series for sufficiently smooth functions $f \in \mathscr{F}_I^{s+1,1}$, where $\mathscr{F}_I^{t,1}$ is a slightly smaller version of $f \in \mathscr{F}^{t,1}$ given by

$$\|f\|_{\mathscr{F}_I^{t,1}} = \int_{\mathbb{R}^d} (1 + \|\xi\|)^t \left|\hat{f}(\xi)\right| d\xi < \infty \tag{11}$$

for $t \in \mathbb{N}$. In theorem 4.2 it is shown that the remainder of the Taylor series of order s is an element of the Barron Space with a higher order ReLU σ_s as activation function. We show that when we approximate this remainder with a shallow neural network, the bound does not suffer from the curse of dimensionality and decreases with the inverse of a factorial of s. The presence of this factorial suggests that higher s give a lower error. We have tested this on a $d = 1$ Gaussian, and have seen that higher s does not give a lower error. The numerical effects dominate and cause the error to increase with increasing s.

The second novel result expands the number of activation functions ϕ for which the Barron spaces can be approximated by Barron spaces with ReLU activation function. Instead of needing

$$\int_{\mathbb{R}} \left|\partial^2\phi(x)\right|(1 + |x|)dx < \infty, \tag{12}$$

it is sufficient that $\phi \in C^2(\mathbb{R})$ or that $\phi = \sigma_s$.

1.3 Structure of this work

After introducing the used notation and various used concepts in section 1.4, we start in section 2 with an explanation of the various error terms we consider in this work. These error terms are based

on the PAC framework. Then, we rigorously define the Barron spaces and related spaces called the Bach spaces in section 3. We show the relation between them in the form of embeddings. In section 3.1 we show that both are Banach spaces. In section 3.2 we use a different formulation of the spaces to expand the set of possible weight functions. With this formulation we establish a link between the Barron spaces and Reproducing Kernel Hilbert Spaces. In section 3.4 we study the effect of changes in the activation function for the Barron and Bach spaces; in particular, we prove the second novel result. In section 3.5 we take the first steps in understanding the dual and predual of Barron space by studying the dual and predual of the Bach spaces.

In section 4 we study the Taylor expansion. We do this by first looking at the Taylor remainder theorem for dimension $d = 1$ in section 4.1. Afterwards, we generalise the Taylor theorem to higher dimensions. In section 4.3 we will go from the multivariate Taylor expansion to the multivariate Taylor expansion in Fourier space. We show that the remainder of this expansion in Fourier space is a Barron function with higher order ReLU as activation function. In section 4.4 we show that the Barron spaces with higher order ReLU as activation function do not suffer from the curse of dimensionality, and provide an error bound for approximating the remainder of this expansion in Fourier space using shallow neural networks.

In section 5 we numerically test the bound derived in section 4.4. This is done using a $d = 1$ Gaussian. In section 5.1 we compute the error bound for the Gaussian analytically. After describing the experiment in more detail in section 5.2, we present and discuss the results in section 5.3.

In section 6 we combine the various elements into a couple of graphs representing embeddings between spaces. In section 6.1 we do this for the Barron spaces with different activation functions, whereas in section 6.2 we do this for $\mathscr{F}_I^{s,1}$ and the Sobolev spaces.

In section 7 we discuss the interplay between deep learning and control. In section 7.1 we show how deep learning helps control by fitting high dimensional functions. We do this from a model adaptive perspective in section 7.1.1, and from an approximate dynamic programming perspective in section 7.1.2. In section 7.2 we study a generalisation of Barron space called the Neural ODE using control methods.

In the final section of this work, section 8, we pose several open questions and conjectures.

1.4 Notation and used concepts

Let $\mathbb{R}$ denote the real numbers, and $\mathbb{N}$ the natural numbers without 0. When we define a function, map or operator f, we write f like

$$f : A \to B, \quad x \mapsto f(x).$$

In this case A and B refer to the input space and output space respectively. Whenever we say f is μ-integrable over C, we mean that $\int_C f(x)d\mu(x)$ exists and is finite. All measurable spaces have the Borel σ-algebra, and evaluating a measure μ on a set A from the sigma algebra means computing

$$\mu(A) = \int_A d\mu(x).$$

If for each Borel set A two measures μ and ν satisfy

$$\mu(A) = 0 \implies \nu(A) = 0,$$

then $\mu \ll \nu$ and there exists a measurable function such that

$$\nu(A) = \int_A f(x) d\mu(x).$$

This is also denoted as

$$d\nu(x) = f(x) d\mu(x).$$

If μ is a signed measure, then there exists a Hahn decomposition such that

$$\mu = \mu_+ - \mu_-$$
$$|\mu| = \mu_+ + \mu_-$$

where $\mu_\pm$ are nonnegative measures. When we say ν is a pushforward of μ along the map Θ given by

$$\Theta : X \to Y, \quad x \mapsto \Theta(x),$$

then we write $\nu := \Theta_\# \mu$ and mean

$$\int_Y f(y) d\nu(y) = \int_X f(\Theta(x)) d\mu(x),$$

where f is μ-measurable. When π is a nonnegative measure with $\pi(X) = 1$, then π is a probability measure. The expectation of f with respect to π is given by

$$\mathbb{E}_\pi[f] = \int_X f(x) d\pi(x).$$

If we sample x from π, then we write $x \sim \pi$. If we sample a set S of n elements from π, then we write $S \sim \pi^n$.

The (strong) derivatives of a function f are represented by $\partial^\alpha f$ and the weak derivatives are represented by $D^\alpha f$. When f is a univariate function $\alpha \in \mathbb{N}$, and when f is a multivariate function α is a multi-index. The gradient of a multivariate function is given by

$$\boldsymbol{\nabla} f = \begin{pmatrix} \partial^{(1,0,\ldots,0)} f \\ \partial^{(0,1,\ldots,0)} f \\ \vdots \\ \partial^{(0,0,\ldots,1)} f \end{pmatrix}.$$

Commonly used sets are $\mathcal{X} \subseteq \mathbb{R}^{d_1}$, $U \subseteq \mathbb{R}^{d_2}$ and $\Omega \subseteq \mathbb{R}^{d_3}$, where $d_i \in \mathbb{N}$. Functions spaces over these sets include

$C^k(\mathcal{X})$ k times continuous differentiable functions,

$C^{0,1}(\mathcal{X})$	bounded Lipschitz continuous functions,
$L^p(\mathcal{X})$	equivalence classes of functions with finite $\|\cdot\|_{L^p(\mathcal{X})}$ norm,
$L^p(\mathcal{X},\mu)$	equivalence classes of functions with finite $\|\cdot\|_{L^p(\mathcal{X},\mu)}$ norm,
$H^k(\mathcal{X})$	Sobolev space of k times weakly differentiable functions with finite $\|\cdot\|_{H^k(\mathcal{X})}$ norm,
$W^{k,p}(\mathcal{X})$	Sobolev space of k times weakly differentiable functions with finite $\|\cdot\|_{W^{k,p}(\mathcal{X},\mathcal{Y})}$ norm,
$\ell^p(\mathcal{X})$	vectors with finite $\|\cdot\|_{\ell^p}$ norm,
$rca(\Omega)$	real, countably additive, signed measures on Σ_Ω with bounded finite variation,
$\mathbb{P}_k(\Omega)$	probability measures on Ω with finite k-th moment.

If we add a 0 subscript, then we mean the same space but restricted to the functions that vanish at infinity when they are defined over $\mathbb{R}^d$ or have zero boundary when defined on a set with a boundary. For example $C_0^k(\mathbb{R}^d)$ consists of k times continuously differentiable functions from $\mathbb{R}^d$ to $\mathbb{R}$ that vanish at infinity.

The norms corresponding to the above spaces are

$$\begin{aligned}
\|f\|_{C^k(\mathcal{X})} &= \sup_{0\leqslant|\alpha|\leqslant k}\sup_{x\in\mathcal{X}}|\partial^\alpha f(x)|,\\
\|f\|_{C^{0,1}(\mathcal{X})} &= \|f\|_{C^0(\mathcal{X})} + \sup_{\substack{x,y\in\mathcal{X}\\ x\neq y}}\frac{|f(x)-f(y)|}{|x-y|},\\
\|f\|_{L^p(\mathcal{X})} &= \left(\int_{\mathcal{X}}|f(x)|^p dx\right)^{1/p},\\
\|f\|_{L^p(\mathcal{X},\mu)} &= \left(\int_{\mathcal{X}}|f(x)|^p d\mu(x)\right)^{1/p},\\
\|f\|_{H^k(\mathcal{X})} &= \sum_{j=0}^{k}\left\|D^j f\right\|_{L^2(\mathcal{X})},\\
\|f\|_{W^{k,p}(\mathcal{X})} &= \sum_{j=0}^{k}\left\|D^j f\right\|_{L^p(\mathcal{X})},\\
\|x\|_{\ell^p(\mathcal{X})} &= (\sum_i |x_i|^p)^{1/p},\\
\|\mu\|_{rca(\Omega)} &= |\mu|(\Omega),\\
\|\pi\|_{\rho_k(\Omega)} &= |\pi|(\Omega) = 1.
\end{aligned}$$

When a space S permits an inner product, then we write $\langle\cdot|\cdot\rangle_S$. If from context no confusion can arise over what inner product we are taking, then we drop the index. Furthermore, we write $\|z\|_p$ instead of $\|z\|_{\ell^p}$ or $\|z\|_{L^p}$, when it is clear whether z is a vector or a function.

We write S instead of $(S,\|\cdot\|_S)$ when referring to the normed vector space S unless context requires us to. Given two normed vector spaces $(S,\|\cdot\|_S)$ and $(T,\|\cdot\|_T)$, then S continuously embeds in T if $S\subseteq T$ and for all $f\in S$ we have $\|f\|_T \leqslant C\|f\|_S$ for some constant $C\in\mathbb{R}$. The symbol for this is $S\hookrightarrow T$. If both $S\hookrightarrow T$ and $T\hookrightarrow S$, then S and T are isomorphic; this is denoted by $\cong$. If additionally $\|f\|_T = \|f\|_S$, then S and T are isometrically isomorphic, which is denoted by $\simeq$.

A function f is called subhomogeneous with order k when for all positive scalar λ it satisfies $f(\lambda x) \leqslant \lambda^k f(x)$, and homogeneous if the inequality holds with equality. A function f is called sigmoidal when it is continuously differentiable, monotone increasing and has limits $\lim_{z\to\pm\infty} f(z) = \pm 1$.

The Fourier transform is a bounded linear operator from $L^1(\mathbb{R}^d) \to L^\infty(\mathbb{R}^d)$ or $L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$. If $f \in L^1(\mathbb{R}^d)$, then its Fourier transform $\hat{f}$ can be computed using

$$\hat{f}(\xi) = \int_{\mathbb{R}^d} f(x) e^{-i\langle x|\xi\rangle} d\xi$$

where

$$c_d = \frac{1}{(2\pi)^{d/2}},$$

and if $\hat{f} \in L^1(\mathbb{R}^d)$, then the inverse operation is given by

$$f(x) = \int_{\mathbb{R}^d} \hat{f}(\xi) e^{i\langle x|\xi\rangle} d\xi.$$

There is no proper integral formula describing the Fourier transform $\hat{f}$, if $f \in L^2(\mathbb{R}^d)$. It can however be computed by considering a sequence of functions $f_n \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ such that

$$f_n \xrightarrow{\|\cdot\|_{L^2(\mathbb{R}^d)}} f.$$

The Fourier transform of $\hat{f}$ then equals the limit of the Fourier transforms $\hat{f}_n$ of the f_n, i.e.

$$\hat{f}_n \xrightarrow{\|\cdot\|_{L^2(\mathbb{R}^d)}} \hat{f}.$$

The Fourier transform is not defined on strict subsets of R^d. To compute the Fourier transform on compact sets U in R^d the functions from $L^1(U)$ or $L^2(U)$ need to be extended to functions from $L^1(\mathbb{R}^d)$ or $L^2(\mathbb{R}^d)$. One way to do so is by using an extension operator $E : L^p(U) \to L^p(\mathbb{R}^d)$ for $p \in \{1, 2\}$, defined by

$$Ef = \begin{cases} f & \text{on } U, \\ 0 & \text{otherwise.} \end{cases}$$

If f has some smoothness properties, e.g. $f \in W^{1,3}(U)$, then this extension will generally not give a function $Ef \in W^{1,3}(\mathbb{R}^d)$. However, if the boundary of U is smooth enough, then for any $p \in \mathbb{N}$ and for any open set O from $\mathbb{R}^d$ such that U is compactly supported in O there exists an extension operator $E : W^{1,p}(U) \to W^{1,p}(\mathbb{R}^d)$ such that for all $f \in W^{1,p}(U)$ $Ef = f$ a.e. on U, Ef is compactly supported within O and there exists a constant C depending on p, d, U and O such that

$$\|Ef\|_{W^{1,p}(\mathbb{R}^d)} \leqslant C\|f\|_{W^{1,p}(U)}.$$

Taking the Fourier transform of $f \in W^{1,3}(U)$ then becomes taking the Fourier transform of $Ef \in W^{1,3}(\mathbb{R}^d)$ for some extension operator E.

An infimum over an empty set is defined as ∞. For $c \in \mathbb{R}$, $N \in \mathbb{N}$, $f : \mathbb{R} \to \mathbb{R}$, and a set A we write

- $c + A = \left\{ c + a \;\middle|\; a \in A \right\}$
- $cA = \left\{ ca \;\middle|\; a \in A \right\}$
- $f(A) = f \circ A = \left\{ f(a) \;\middle|\; a \in A \right\}$
- $\operatorname{conv} A = \left\{ \sum_{i=1}^{N} c_i a_i \;\middle|\; a_i \in A, \sum_{i=1}^{N} c_i = 1, c_i \geqslant 0 \right\}$

Finally, consider a subset U of a normed vector space, then R_U denotes the (possibly infinite) radius of the smallest closed ball centered at the origin that fully contains U.

2 Framework for Estimating functions using Neural Networks

In this work we discuss approximating a map $f^\star$ using neural networks. Approximating functions leads to error terms. The errors we are considering are listed in fig. 1. These errors are similar to those defined in the Probably Approximately Correct (PAC) framework[Mohri et al., 2012; Shalev-Shwartz and Ben-David, 2014]. In this section we will define these errors more rigorously, and we will discuss the methods that are used later to estimate them.

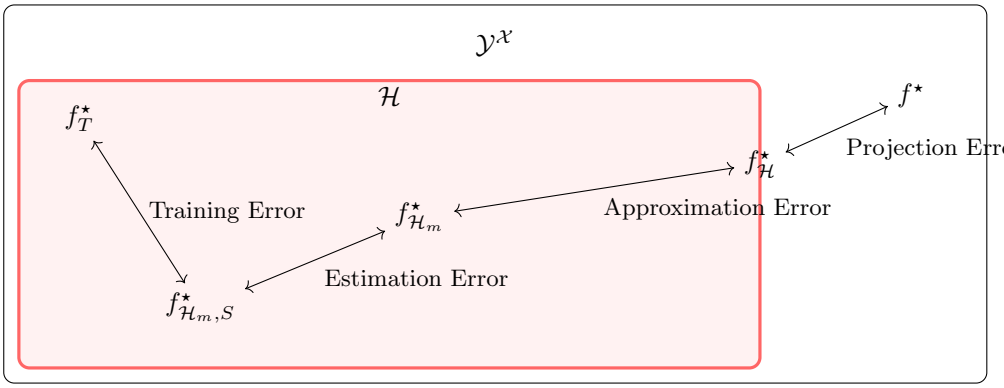


Figure 1: Visual representation of the various errors. $\mathcal{Y}^{\mathcal{X}}$ is the set of all functions from $\mathcal{X}$ to $\mathcal{Y}$, and the hypothesis space $\mathcal{H}$ is the set of 'nice' functions we are considering. The projection error is the error between the true data f^* and $f^*_{\mathcal{H}}$, the best fit within $\mathcal{H}$. The approximation error is the error between $f^*_{\mathcal{H}}$ and the best fit within $\mathcal{H}$ of m elements, $f^*_{\mathcal{H}_m}$. The estimation error is the error between $f^*_{\mathcal{H}_m}$ and the best fit within $\mathcal{H}$ of m elements for $S \subset \mathcal{X}$, $f^*_{\mathcal{H}_m,S}$. Lastly, the training error is the error between $f^*_{\mathcal{H}_m,S}$ and what is found using training, $f^\star_T$.

2.1 Problem Description

Consider an input space $\mathcal{X}$, a target space $\mathcal{Y}$, a map $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty]$, and a probability measure $\rho \in \mathbb{P}_k(\mathcal{X})$ for some $k \in \mathbb{N}$. Let

$$\mathcal{Y}^{\mathcal{X}} = \{f \mid f : \mathcal{X} \to \mathcal{Y}\}, \tag{13}$$

i.e. $\mathcal{Y}^{\mathcal{X}}$ is the set of all the functions from $\mathcal{X}$ to $\mathcal{Y}$. Lastly, define the loss functional

$$\mathcal{L} : \mathcal{Y}^{\mathcal{X}} \times \mathcal{Y}^{\mathcal{X}} \to [0, \infty), \ (f, g) \mapsto \int_{\mathcal{X}} \ell(f(x), g(x)) d\rho(x) \tag{14}$$

with the loss function

$$\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty). \tag{15}$$

The probability measure π represents the way the data is sampled from $\mathcal{X}$, the function ℓ represents a pointwise penalty, and the loss $\mathcal{L}$ aggregates the penalty over the sampled data.

We are interested in approximating some function $f^\star$. We use the loss functional $\mathcal{L}$ to determine how close our approximation is to the function of interest $f^\star$. Logically, we want ℓ to be such that

$$\mathcal{L}(f^\star, f^\star) = 0. \tag{16}$$

Examples of functions ℓ used in practice that achieve this are

$$\ell(u, v) = (u - v)^2 \tag{17}$$

and

$$\ell(u, v) = \begin{cases} 0 & u = v, \\ 1 & u \neq v. \end{cases} \tag{18}$$

From eq. (16) is follows that $f^\star$ satisfies

$$f^\star = \arg\min_{f \in \mathcal{Y}^{\mathcal{X}}} \mathcal{L}(f, f^\star). \tag{19}$$

Equation (19) is a recursive problem. In order to find the minimum argument $f^\star \in \mathcal{Y}^{\mathcal{X}}$, we need to choose an ℓ and know $f^\star$ in order to evaluate $\mathcal{L}(f, f^\star)$. Even if we did not know $f^\star$ and did have an oracle

$$\mathcal{O} : \mathcal{X} \times \mathcal{Y}^{\mathcal{X}} \to [0, \infty), \ (x, f) \mapsto \ell(f(x), f^\star(x)) \tag{20}$$

that tells us the pointwise penalty at x for a given function f, we cannot solve eq. (19) in practice. This is due to four reasons:

1. $\mathcal{Y}^{\mathcal{X}}$ is typically too large to search through.
2. We can only use a finite number of parameters to describe the function f.
3. We only have access to a finite number of samples from ρ with the corresponding $f^\star$ values.
4. We have to use an algorithm to look for the best f, and this does not have to yield an optimal solution.

The first of these can be addressed by restricting the set over which is optimised. To restrict $\mathcal{Y}^{\mathcal{X}}$ we consider a hypothesis space $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$. The best map inside the hypothesis space $\mathcal{H}$ that we are able to find will be

$$f^\star_{\mathcal{H}} = \arg\min_{f \in \mathcal{H}} \mathcal{L}(f, f^\star). \tag{21}$$

Ideally this hypothesis space is large enough to cover most functions of interest in $\mathcal{Y}^{\mathcal{X}}$ whilst simultaneously allowing us to easily solve eq. (21). These two properties are trade-offs; making the space $\mathcal{H}$ bigger typically means that eq. (21) is harder to solve. It heavily depends on the problem what a suitable choice for $\mathcal{H}$ is. Regardless of the choice of hypothesis space $\mathcal{H}$, we do not expect

$$\mathcal{L}(f^\star_{\mathcal{H}}, f^\star) = 0 \tag{22}$$

to hold in general. This quantity $\mathcal{L}(f^\star_{\mathcal{H}}, f^\star)$ is called the *projection error*.

The second reason can be addressed by looking only at a subset of $\mathcal{H}$. Functions in $\mathcal{H}$ potentially have infinitely many parameters. Let $\mathcal{H}_m$ contain the functions from $\mathcal{H}$ that at most require m parameters. The best map inside it that we will be able to find is

$$f^\star_{\mathcal{H}_m} = \arg\min_{f \in \mathcal{H}_m} \mathcal{L}(f, f^\star). \tag{23}$$

Again, we don't expect

$$\mathcal{L}(f^\star_{\mathcal{H}_m}, f^\star) = \mathcal{L}(f^\star_{\mathcal{H}}, f^\star) \tag{24}$$

to hold in general. The quantity $\mathcal{L}(f^\star_{\mathcal{H}_m}, f^\star) - \mathcal{L}(f^\star_{\mathcal{H}}, f^\star)$ is called the *approximation error*, and it is a quantifier for the expressivity of the hypothesis space.

The third reason effectively states that we cannot evaluate $\mathcal{L}$ at (f, g), but we can only evaluate the loss functional $\mathcal{L}_S$ defined by

$$\mathcal{L}_S : \mathcal{Y}^{\mathcal{X}} \times \mathcal{Y}^{\mathcal{X}} \to [0, \infty),\ (f, g) \mapsto \frac{1}{|S|} \sum_{x_i \in S} \ell(f(x_i), g(x_i)), \tag{25}$$

where $S = (x_i)_{i=1}^m$ represents $m \in \mathbb{N}$ i.i.d samples from ρ, at (f, g). S is called the *training set*.

Remark. *Although S is called the training set, it is not a set. It is a sequence of $|S|$ tuples. This allows for an ordering and non-unique elements. This makes a difference for the algorithms used to solve the minimisation problem.*

This sampling introduces another error, commonly called the *estimation error* or *sample error*, and is a quantifier for how well $\hat{f}_{\mathcal{H}_m}$ can be reconstructed from data. The error is given by

$$\mathcal{L}(f^\star, f^\star_{\mathcal{H}_m,S}) - \mathcal{L}(f^\star, f^\star_{\mathcal{H}_m}), \tag{26}$$

where

$$f^\star_{\mathcal{H}_m,S} = \arg\min_{f \in \mathcal{H}_m} \mathcal{L}_S(f, f^\star). \tag{27}$$

The fourth reason also gives an error. This final error is due to the spaces $\mathcal{H}$ and $\mathcal{H}_m$ usually being non-convex. Solving the resulting numerical non-convex optimisation problem in finite time can usually only be done approximately. This process is called *training*. Denote with $f^\star_T$ the map retrieved by training; the *training error* is then given by

$$\mathcal{L}(f^\star, f^\star_T) - \mathcal{L}(f^\star, f^\star_{\mathcal{H}_m,S}).$$

So, whilst we wish to find $f^\star$, we are only compute $f^\star_T$ in practice. Only in the most trivial cases will we have that $f^\star_T$ and $f^\star$ match. We are interested in knowing how much they differ, i.e. we want to know $\mathcal{L}(f^\star, f^\star_T)$. Using the four discussed errors we can write

$$\begin{aligned}\mathcal{L}(f^\star, f^\star_T) = &\underbrace{\mathcal{L}(f^\star, f^\star_{\mathcal{H}})}_{\text{projection error}} + \underbrace{\left(\mathcal{L}(f^\star, f^\star_{\mathcal{H}_m}) - \mathcal{L}(f^\star, f^\star_{\mathcal{H}})\right)}_{\text{approximation error}} \\ &+ \underbrace{\left(\mathcal{L}(f^\star, f^\star_{\mathcal{H}_m,S}) - \mathcal{L}(f^\star, f^\star_{\mathcal{H}_m})\right)}_{\text{estimation error}} + \underbrace{\left(\mathcal{L}(f^\star, f^\star_T) - \mathcal{L}(f^\star, f^\star_{\mathcal{H}_m,S})\right)}_{\text{training error}}.\end{aligned} \tag{28}$$

Each of the four error terms is heavily influenced by the choice of hypothesis space $\mathcal{H}$ and ℓ. Hence, if we want to find good approximations $f_T^\star$ of $f^\star$ based on finite data samples S using only m parameters, then we should study how the four discussed error terms depend on the hypothesis space $\mathcal{H}$ and ℓ. Doing this in general is outside the scope of this thesis. In this work we will focus on the Barron spaces as hypothesis spaces and only consider

$$\ell(u, v) = (u - v)^2 \tag{29}$$

such that

$$\mathcal{L}(f, g) = \|f - g\|^2_{L^2(\mathcal{X},\rho)}. \tag{30}$$

In the remainder of this section we will focus on methods for bounding the four errors.

2.2 Error Bounds

We have discussed the projection error, approximation error, estimation error and the training error. For the approximation error and estimation error various concepts and methods are available to bound them. However, there are no general methods available for bounding the projection error and the training error. Hence, we will consider the latter two out of the scope of this thesis. We will now show one method for bounding the approximation error as well as one method for bounding the estimation error. The method for bounding the estimation error relies on [Mohri et al., 2012; Shalev-Shwartz and Ben-David, 2014]. Most of the proofs can be found there too, but they have been included for completeness and adapted to the notation of this work.

The approximation error seems to depend on $f^\star$, but is in fact independent of $f^\star$ for our choice of ℓ, as is shown in proposition 2.1.

Proposition 2.1.

$$\mathcal{L}(f^\star, f^\star_{\mathcal{H}_m}) - \mathcal{L}(f^\star, f^\star_{\mathcal{H}}) = \mathcal{L}(f^\star_{\mathcal{H}}, f^\star_{\mathcal{H}_m})$$

Proof.

$$\begin{aligned}
\mathcal{L}(f^\star, f^\star_{\mathcal{H}_m}) &= \left\|f^\star - f^\star_{\mathcal{H},m}\right\|^2_{L^2(\mathcal{X},\rho)} \\
&= \left\|(f^\star - f^\star_{\mathcal{H}}) + (f^\star_{\mathcal{H}} - f^\star_{\mathcal{H},m})\right\|^2_{L^2(\mathcal{X},\rho)} \\
&= \|f^\star - f^\star_{\mathcal{H}}\|^2_{L^2(\mathcal{X},\rho)} + \left\|f^\star_{\mathcal{H}} - f^\star_{\mathcal{H},m}\right\|^2_{L^2(\mathcal{X},\rho)} + 2\left\langle f^\star - f^\star_{\mathcal{H}} \middle| f^\star_{\mathcal{H}} - f^\star_{\mathcal{H},m}\right\rangle_{L^2(\mathcal{X},\rho)} \\
&= \mathcal{L}(f^\star, f^\star_{\mathcal{H}}) + \mathcal{L}(f^\star_{\mathcal{H}}, f^\star_{\mathcal{H}_m}) + 2\left\langle f^\star - f^\star_{\mathcal{H}} \middle| f^\star_{\mathcal{H}} - f^\star_{\mathcal{H},m}\right\rangle_{L^2(\mathcal{X},\rho)}
\end{aligned}$$

By definition of $f^\star_{\mathcal{H}}$

$$\left\langle f^\star - f^\star_{\mathcal{H}} | g\right\rangle_{L^2(\mathcal{X},\rho)} = 0$$

for all $g \in \mathcal{H}$. Since $f^\star_{\mathcal{H}} - f^\star_{\mathcal{H},m} \in \mathcal{H}$,

$$\left\langle f^\star - f^\star_{\mathcal{H}} \middle| f^\star_{\mathcal{H}} - f^\star_{\mathcal{H},m}\right\rangle_{L^2(\mathcal{X},\rho)} = 0.$$

Hence,

$$\mathcal{L}(f^\star, f^\star_{\mathcal{H}_m}) = \mathcal{L}(f^\star, f^\star_{\mathcal{H}}) + \mathcal{L}(f^\star_{\mathcal{H}}, f^\star_{\mathcal{H}_m}).$$

Rearranging finishes the proof. *Q.E.D.*

Proposition 2.1 shows that the approximation error depends on the relation between $\mathcal{H}$ and $\mathcal{H}_m$. Hence, we cannot refine this further without specifying $\mathcal{H}$.

For the estimation error we have $f^\star_{\mathcal{H}_m,S}$. This function depends on S, but S is randomly sampled. This means that we expect an upper bound for the estimation error to be probabilistic. To formulate this we will use the Rademacher complexity and the representativeness.

***Definition* 1** (Representativeness)**.** *Let $\rho \in \mathbb{P}(\mathbb{R}^d)$ be a probability measure, $\mathcal{F}$ be a set of functions for which $\mathbb{E}_\rho[f]$ is well defined and finite when $f \in \mathcal{F}$, and S a collection of n points sampled from ρ. The quantity*

$$\operatorname*{Rep}_S(\mathcal{F}) = \sup_{f\in\mathcal{F}} \mathbb{E}_\rho[f] - \frac{1}{|S|}\sum_{x\in S} f(x)$$

is called the representativeness.

A low representativeness means that for each $f \in \mathcal{F}$ the average over S is similar to the expectation over ρ. At the same time we know from the definition of $f_{\mathcal{H}_m,S}$ that

$$\mathcal{L}_S(f^\star, f^\star_{\mathcal{H}_m}) \geqslant \mathcal{L}_S(f^\star, f^\star_{\mathcal{H}_m,S}). \tag{31}$$

We can use this to bound the estimation error by the representativeness:

$$\begin{aligned}
\mathcal{L}(f^\star, f^\star_{\mathcal{H}_m,S}) - \mathcal{L}(f^\star, f^\star_{\mathcal{H}_m}) &= \mathcal{L}(f^\star, f^\star_{\mathcal{H}_m,S}) - \mathcal{L}_S(f^\star, f^\star_{\mathcal{H}_m,S}) + \mathcal{L}_S(f^\star, f^\star_{\mathcal{H}_m,S}) - \mathcal{L}(f^\star, f^\star_{\mathcal{H}_m}) \\
&\leqslant \mathcal{L}(f^\star, f^\star_{\mathcal{H}_m,S}) - \mathcal{L}_S(f^\star, f^\star_{\mathcal{H}_m,S}) + \mathcal{L}_S(f^\star, f^\star_{\mathcal{H}_m}) - \mathcal{L}(f^\star, f^\star_{\mathcal{H}_m}) \\
&\leqslant 2\sup_{f\in\mathcal{H}}(\mathcal{L}(f^\star, f) - \mathcal{L}_S(f^\star, f)) \\
&= 2\operatorname*{Rep}_S(\mathcal{F}),
\end{aligned} \tag{32}$$

where

$$\mathcal{F} = \left\{ x \mapsto \ell(f(x), f^\star(x)) \;\middle|\; f \in \mathcal{H} \right\}. \tag{33}$$

To compute the representativeness of $\mathcal{F}$ for a set S we still need to know ρ. If $\mathcal{F}$ has a low representativeness for S_1, then the representativeness for another set S_2 of the same size n is roughly equal to

$$\frac{1}{n}\sup_{f\in\mathcal{F}} \sum_{x\in S_1} f(x) - \sum_{x\in S_2} f(x). \tag{34}$$

This can be written more compactly as

$$\sup_{f\in\mathcal{F}} \sum_{i=1}^{|S|} \chi_i f(x_i) \tag{35}$$

by considering

$$\chi_i = \begin{cases} 1 & i \leqslant n \\ -1 & i > n \end{cases} \tag{36}$$

and concatenating S_2 to S_1 to get a single training set S, i.e. the first n elements of S are those from S_1 and the following n elements are those from S_2. Instead of picking two sets S_1 and S_2 and combining this into one set S, we can also start with a training set S and split it into two sets S_1 and S_2 by doing coin flips. Head means $x \in S$ becomes $x \in S_1$ and tails means $x \in S_2$. The expected outcome after doing coin flips is the Rademacher complexity.

***Definition* 2** (Rademacher complexity)**.** *The empirical Rademacher complexity is given by*

$$\operatorname*{Rad}_S(\mathcal{F}) = \frac{1}{|S|}\mathbb{E}_\chi\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{|S|}\chi_i f(z_i)\right], \tag{37}$$

where χ_i are i.i.d. random variables with $\mathbb{P}(\chi_i = 1) = \mathbb{P}(\chi_i = -1) = \frac{1}{2}$. Its expectation

$$\operatorname{Rad}(\mathcal{F}; n) = \mathbb{E}_{S\sim\rho^n}\operatorname*{Rad}_S(\mathcal{F}) \tag{38}$$

is the Rademacher complexity.

Intuitively, the Rademacher complexity quantifies how well functions from $\mathcal{F}$ can be used to fit random noise. Sets of functions with higher Rademacher complexity tend to be able to fit more complex functions. Note that the Rademacher complexity is an expectation, so it no longer depends on the chosen S but only on the properties of $\mathcal{F}$. If we can bound the representativeness using the Rademacher complexity, we have gotten rid of the dependence on S of our upper bound for the approximation error. The link between the Rademacher complexity and the expectation of the representativeness for a function is given in proposition 2.2.

Proposition 2.2. *Let $\pi \in \mathbb{P}(\mathbb{R}^d)$ be a probability measure, and $\mathcal{F}$ be a set of functions for which $\mathbb{E}_\pi[f]$ is well defined and finite when $f \in \mathcal{F}$, then*

$$\mathbb{E}_{S\sim\pi^n}[\operatorname*{Rep}_S(\mathcal{F})] \leqslant 2\operatorname{Rad}(\mathcal{F}; n). \tag{39}$$

Proof. This follows from the sequence of (in)equalities

$$\begin{aligned}
\mathbb{E}_{S\sim\pi^n}[\operatorname*{Rep}_S(\mathcal{F})] &= \mathbb{E}_{S\sim\pi^n}[\sup_{f\in\mathcal{F}}\mathbb{E}_\pi[f] - \frac{1}{|S|}\sum_{x\in S}f(x)] \\
&= \mathbb{E}_{S\sim\pi^n}[\sup_{f\in\mathcal{F}}\mathbb{E}_{\tilde{S}\sim\pi^n}[\frac{1}{\left|\tilde{S}\right|}\sum_{x\in\tilde{S}}f(x)] - \frac{1}{|S|}\sum_{x\in S}f(x)] \\
&= \mathbb{E}_{S\sim\pi^n}[\sup_{f\in\mathcal{F}}\mathbb{E}_{\tilde{S}\sim\pi^n}[\frac{1}{n}\sum_{x\in\tilde{S}}f(x)] - \frac{1}{n}\sum_{x\in S}f(x)] \\
&= \mathbb{E}_{S\sim\pi^n}[\sup_{f\in\mathcal{F}}\mathbb{E}_{\tilde{S}\sim\pi^n}[\frac{1}{n}\sum_{(x,y)\in\tilde{S}\times S}f(x) - f(y)]]
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \mathbb{E}_{S\sim\pi^n}[\mathbb{E}_{\tilde{S}\sim\pi^n}[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{(x,y)\in\tilde{S}\times S} f(x)-f(y)]] && \text{Jensen's ineq.}\\
&= \mathbb{E}_{\chi}\mathbb{E}_{S\sim\pi^n}[\mathbb{E}_{\tilde{S}\sim\pi^n}[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{(x,y)\in\tilde{S}\times S} \chi(f(x)-f(y))]] && \mathcal{F}\text{ symmetric}\\
&= \mathbb{E}_{\chi}\mathbb{E}_{S\sim\pi^n}[\mathbb{E}_{\tilde{S}\sim\pi^n}[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{x\in\tilde{S}}\chi f(x)+\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{y\in S}-\chi f(y)]]\\
&= \mathbb{E}_{\chi}\mathbb{E}_{\tilde{S}\sim\pi^n}[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{x\in\tilde{S}}\chi f(x)]+\mathbb{E}_{\chi}\mathbb{E}_{S\sim\pi^n}[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{y\in S}-\chi f(y)]\\
&= \mathbb{E}_{\chi}\mathbb{E}_{\tilde{S}\sim\pi^n}[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{x\in\tilde{S}}\chi f(x)]+\mathbb{E}_{\chi}\mathbb{E}_{S\sim\pi^n}[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{y\in S}\chi f(y)] && \mathcal{F}\text{ symmetric}\\
&= 2\,\mathrm{Rad}\,(\mathcal{F};n)
\end{aligned}
$$

Q.E.D.

This is a bound for the expectation of the representativeness in terms of the Rademacher complexity, not a bound for the representativeness itself. For that we use *McDiarmid's inequality.*

Lemma 2.0.1 (McDiarmid's inequality)**.** *Let $V\subseteq\mathbb{R}^d$ be some set and let $g: V^n\to\mathbb{R}$ be a function of n variables such that for some $c>0$, for all $i\in[n]$ and for all $x_1,\dots,x_n,x_i'\in V$ we have*

$$|g(x_1,\dots,x_n)-g(x_1,\dots,x_{i-1},x_i',x_{i+1},\dots,x_n)|\leqslant c. \tag{40}$$

Let X_i for $i\in[1,\dots,n]$ be i.i.d. random variables taking values in V, then

$$\mathbb{P}\left(\left|g(X_1,\dots,X_n)-\mathbb{E}[g(X_1,\dots,X_n)]\right|\leqslant c\sqrt{\frac{n}{2}\log\left(\frac{2}{\delta}\right)}\right)\geqslant 1-\delta. \tag{41}$$

McDiarmid's inequality allows us to establish a bound for the representativeness in terms of the Rademacher complexity that holds with high probability, and thus a bound for the estimation error.

Proposition 2.3. *If the functions in $\mathcal{F}$ are bounded by M and L-Lipschitz with respect to $\mathcal{H}$, then with probability at least $1-\delta$*

$$\mathcal{L}(f^\star,f^\star_{\mathcal{H}_m,S})-\mathcal{L}(f^\star,f^\star_{\mathcal{H}_m})\leqslant 4L\,\mathrm{Rad}\,(\mathcal{H};n)+2M\sqrt{\frac{2\log\left(\frac{2}{\delta}\right)}{n}} \tag{42}$$

over the sets S with $|S|=n$.

Proof. Let $V=\operatorname{supp}\rho$,

$$g: V^n\to\mathbb{R},\ S\mapsto\operatorname*{Rep}_S(\mathcal{F}), \tag{43}$$

and

$$S_1=\left\{z_1,\dots,z_j,\dots z_n\right\}$$

$$S_2 = \left\{ z_1, \dots, z_j', \dots z_n \right\}$$

with $z_1, ..., z_n, z_j' \in V$. We will show that

$$|g(S_1) - g(S_2)| \leqslant \frac{2M}{n}. \tag{44}$$

To achieve that set

$$f_{S_i} = \arg\max_{f \in \mathcal{F}} \left(\mathbb{E}_\rho[f] - \frac{1}{n} \sum_{z_j \in S_i} f(z_j) \right), \tag{45}$$

and observe that

$$\operatorname*{Rep}_{S_i}(\mathcal{F}) = \mathbb{E}_\rho[f_{S_i}] - \frac{1}{n} \sum_{z_j \in S_i} f_{S_i}(z_j) \tag{46}$$

and

$$\mathbb{E}_\rho[f_{S_1}] - \frac{1}{n} \sum_{z_j \in S_2} f_{S_1}(z_j) \leqslant \mathbb{E}_\rho[f_{S_2}] - \frac{1}{n} \sum_{z_j \in S_2} f_{S_2}(z_j). \tag{47}$$

This implies that

$$\begin{aligned}
|g(S_1) - g(S_2)| &= \left| \left(\mathbb{E}_\rho[f_{S_1}] - \frac{1}{n} \sum_{z_j \in S_1} f_{S_1}(z_j) \right) - \left(\mathbb{E}_\rho[f_{S_2}] - \frac{1}{n} \sum_{z_j \in S_2} f_{S_2}(z_j) \right) \right| \\
&\leqslant \left| \left(\mathbb{E}_\rho[f_{S_1}] - \frac{1}{n} \sum_{z_j \in S_1} f_{S_1}(z_j) \right) - \left(\mathbb{E}_\rho[f_{S_1}] - \frac{1}{n} \sum_{z_j \in S_2} f_{S_1}(z_j) \right) \right| \\
&= \left| \frac{1}{n} \sum_{z_j \in S_1} f_{S_1}(z_j) - \frac{1}{n} \sum_{z_j \in S_2} f_{S_1}(z_j) \right| \\
&= \frac{1}{n} \left| \sum_{z_i \in S_2} f_{S_1}(z_i) - \sum_{z_i \in S_1} f_{S_1}(z_i) \right|.
\end{aligned}$$

By construction S_1 and S_2 differ only in one element, thus

$$\begin{aligned}
|g(S_1) - g(S_2)| &\leqslant \frac{1}{n} \left| \sum_{z_i \in S_2} f_{S_1}(z_i) - \sum_{z_i \in S_1} f_{S_1}(z_i) \right| \\
&\leqslant \frac{1}{n} \left| f_{S_1}(z_j') - f_{S_1}(z_j) \right|.
\end{aligned}$$

By assumption of the functions in $\mathcal{F}$ being bounded by M it follows that

$$\left| f_{S_1}(z_j') - f_{S_1}(z_j) \right| \leqslant 2M. \tag{48}$$

Therefore

$$\left| \operatorname*{Rep}_{S_1}(\mathcal{F}) - \operatorname*{Rep}_{S_2}(\mathcal{F}) \right| = |g(S_1) - g(S_2)| \leqslant \frac{2M}{n}. \tag{49}$$

Applying McDiarmid's inequality to g gives

$$\mathbb{P}\left(\left|\operatorname*{Rep}_{S}(\mathcal{F}) - \mathbb{E}_{S'}[\operatorname*{Rep}_{S'}(\mathcal{F})]\right| \leqslant \frac{2M}{n}\sqrt{\frac{n\log\left(\frac{2}{\delta}\right)}{2}}\right) \geqslant 1-\delta. \tag{50}$$

where the sets S, S' are of size $|S| = |S'| = n$. Hence, with probability at least $1-\delta$

$$\operatorname*{Rep}_{S}(\mathcal{F}) \leqslant \mathbb{E}_{S'}[\operatorname*{Rep}_{S'}(\mathcal{F})] + M\sqrt{\frac{2\log\left(\frac{2}{\delta}\right)}{n}}. \tag{51}$$

By using the bound for the representativeness of proposition 2.2, we have

$$\operatorname*{Rep}_{S}(\mathcal{F}) \leqslant 2\,\mathrm{Rad}\,(\mathcal{F}; n) + M\sqrt{\frac{2\log\left(\frac{2}{\delta}\right)}{n}} \tag{52}$$

with probability at least $1-\delta$. Combining eq. (32) with eq. (52) gives a bound for the estimation error in terms of $\mathscr{F}$.

$$\mathcal{L}(f^\star, f^\star_{\mathcal{H}_m,S}) - \mathcal{L}(f^\star, f^\star_{\mathcal{H}_m}) \leqslant 4\,\mathrm{Rad}\,(\mathcal{F}; m) + 2M\sqrt{\frac{2\log\left(\frac{2}{\delta}\right)}{n}} \tag{53}$$

Finally, according to [Wolf, 2018; theorem 1.14 and lemma 2.7] it holds that

$$\mathrm{Rad}\,(\mathcal{F}; n) \leqslant L\,\mathrm{Rad}\,(\mathcal{H}; n) \tag{54}$$

by the assumptions on the functions in $\mathcal{F}$. Substituting eq. (54) into eq. (53) allows us to rewrite the bound of estimation error in terms of $\mathcal{H}$ and not in terms of $\mathcal{F}$. *Q.E.D.*

3 Infinitely Wide Neural Network Spaces

In the previous section the optimisation problem of interest has been established, and the need for a properly chosen hypothesis space is explained. In this section the optimisation problem will be further specified. In particular, the Bach and Barron spaces will be considered as hypothesis spaces.

Fix $d \in \mathbb{N}$. Let $(\mathcal{X}, (x,y) \mapsto \|x-y\|_{\ell^\infty})$ be a metric space with its metric induced by the ℓ^∞ norm defined over a set $\mathcal{X} \subset \mathbb{R}^d$ with $\mathcal{X}$ a compact set and the boundary of $\mathcal{X}$ smooth. Let $(\Omega, (x,y) \mapsto \|x-y\|_{\ell^1})$ be a metric space with its metric induced by the ℓ^1 norm defined over a set $\Omega \subset \mathbb{R}^{d+1}$ with Ω a non-empty closed set. Restrict the possible sets for Ω and $\mathcal{X}$ to sets that contain a closed ball centered at the origin with positive radius. Let $\mathcal{Y}$ be the real numbers, and ρ be any probability measure with full support on $\mathcal{X}$. $\mathcal{X}$ and $\mathcal{Y}$ represent the input and output spaces for the Barron and Bach spaces. Ω represents the parameters that can be used to construct functions in these spaces. When we write $(w,b) \in \Omega$, w refers to the first d coordinates and b only to the last coordinate. To construct a neural network with parameters in Ω, we use the construction operator.

***Definition* 3** (Activation function; construction operator)**.** *If $\phi : \mathbb{R} \to \mathbb{R}$ is a nonzero function that is applied pointwise to vectors, i.e.*

$$\phi(z) = \begin{pmatrix}\phi(z_1) & \dots & \phi(z_d)\end{pmatrix}^\intercal \tag{55}$$

for all $z \in \mathbb{R}^d$, then ϕ is called an activation function, and the operator

$$K_\phi^\Omega : rca(\Omega) \to (\mathcal{X} \to \mathbb{R}), \quad \mu \mapsto \left(x \mapsto \int_\Omega \phi(\langle x|w\rangle + b) d\mu(w,b) \right) \tag{56}$$

is called the construction operator.

The construction operator turns a measure $\mu \in rca(\Omega)$ into an (infinitely wide) shallow neural network $K_\phi^\Omega \mu$. In particular, if μ is a sum of m Dirac measures, i.e.

$$\mu = \sum_{i=1}^m a_i \delta_{(w_i,b_i)} \tag{57}$$

for $(w_i, b_i) \in \Omega$ and $a_i \in \mathbb{R}$, then

$$K_\phi^\Omega \mu : \mathcal{X} \to \mathbb{R},\ x \mapsto \sum_{i=1}^m a_i \phi(\langle x|w_i\rangle + b_i) \tag{58}$$

is a shallow neural network. For different $\mu \in rca(\Omega)$ it is possible that $K_\phi^\Omega \mu$ describes the same function $f : \mathcal{X} \to \mathbb{R}$. We group those together into $M_{\phi,f}^\Omega$ given by

$$M_{\phi,f}^\Omega = \left\{ \mu \in rca(\Omega) \;\middle|\; \forall x \in \mathcal{X} :\ K_\phi^\Omega \mu(x) = f(x) \right\}. \tag{59}$$

The special case for $f(x) = 0$ is denoted by $M_{\phi,0}^\Omega$. The Barron and Bach spaces are determined by those functions for which $M_{\phi,f}^\Omega$ is non-empty.

***Definition* 4.** *Let $W : \Omega \to \mathbb{R}$ be a non-negative function, and let ϕ be an activation function. The space*

$$\begin{aligned} \mathcal{N}_{\phi,W}^\Omega &= \left\{ f : \mathcal{X} \to \mathbb{R} \;\middle|\; \|f\|_{\mathcal{N}_{\phi,W}^\Omega} < \infty \right\} \\ \|f\|_{\mathcal{N}_{\phi,W}^\Omega} &= \inf_{\mu \in M_f^\phi} \|\mu\|_{W,\Omega} \\ \|\mu\|_{W,\Omega} &= \int_\Omega W(\omega) d|\mu|(\omega) \end{aligned} \tag{60}$$

is called an infinitely wide neural network space. If

$$W(w,b) = 1, \tag{61}$$

then $(\mathcal{N}_{\phi,W}^\Omega, \|\cdot\|_{\mathcal{N}_{\phi,W}^\Omega})$ is denoted $(V_\phi^\Omega, \|\cdot\|_{V_\phi^\Omega})$. If ϕ is such that

$$W_\phi : \Omega \to \mathbb{R},\ (w,b) \mapsto \begin{cases} (\|w\|_1 + |b|)^\alpha & \phi(x) = \sigma_\alpha := \max(0,x)^\alpha, \alpha \in \mathbb{N} \\ 1 + \|w\|_1 + |b| & \phi \text{ Lipschitz} \\ 1 + \|w\|_1 & \phi \in C^{(0,1)}(\mathbb{R}) \end{cases} \tag{62}$$

is well defined, then $(\mathcal{N}_{\phi,W_\phi}^\Omega, \|\cdot\|_{\mathcal{N}_{\phi,W_\phi}^\Omega})$ is denoted $(\mathcal{B}_\phi^\Omega, \|\cdot\|_{\mathcal{B}_\phi^\Omega})$. V_ϕ^Ω is called a Bach space, and $\mathcal{B}_\phi^\Omega$ is called a Barron space.

Remark. *In literature* $(\mathcal{F}_1, \gamma_1)$ *is written instead of* $(V_\phi^\Omega, \|\cdot\|_{V_\phi^\Omega})$*[Bach, 2017]. The notation was changed, because in this thesis the Fourier transform is denoted by* $\mathcal{F}$ *as well as to add explicit dependence on the activation function* ϕ *and the set* Ω*.*

Remark. *Definition A.2 of [E and Wojtowytsch, 2020a] uses a slightly different definition of the Barron spaces. The authors use the ReLU* σ_1 *instead of the higher order ReLU* σ_α *we use here. In [E and Wojtowytsch, 2020b] there are 6 other formulations for the Barron space. Two of these will appear in later sections.*

Remark. *Although the Barron spaces are not defined for many activation functions, this definition covers all commonly used activation functions.*

From definition 4 it follows that the Bach and Barron spaces only differ in the weights they assign to $(w, b) \in \Omega$ in their norms. The Bach spaces do not differentiate between the various $(w, b) \in \Omega$. This makes them the simpler function spaces to analyse, but in practice smaller weights w and biases b are preferred over larger ones. The Barron spaces are an adaptation of the Bach spaces that take into account these weights and biases for (locally) Lipschitz continuous activation functions. This ensures the Barron spaces are more realistic, but also harder to analyse. However, in many cases this change in weight function can be bounded by multiplying the other norm with a constant.

Proposition 3.1. *Let* ϕ *be an activation function for which* W_ϕ *from eq.* (62) *is well defined, but not a higher order ReLU* σ_α*. If* $\Omega \subseteq \mathbb{R}^{d+1}$ *is not a compact set, then* $\mathcal{B}_\phi^\Omega \hookrightarrow V_\phi^\Omega$*. If* $\Omega \subseteq \mathbb{R}^{d+1}$ *is a compact set, then* $V_\phi^\Omega \cong \mathcal{B}_\phi^\Omega$*.*

Proof. The assumption on ϕ implies that

$$W(w, b) = 1 \leqslant W_\phi(w, b) \tag{63}$$

for all $(w, b) \in \Omega$. If $\mu \in M_{\phi,f}^\Omega$ for $f \in \mathcal{B}_\phi^\Omega$, then

$$\|f\|_{V_\phi^\Omega} \leqslant \int_\Omega d|\mu|(w, b) \leqslant \|\mu\|_{W_\phi,\Omega}. \tag{64}$$

Taking the infimum over $\mu \in M_{\phi,f}^\Omega$ gives

$$\|f\|_{V_\phi^\Omega} \leqslant \|f\|_{\mathcal{B}_\phi^\Omega}. \tag{65}$$

This shows the first statement.

If Ω is a compact set, then it is closed and bounded by Heine-Borel. Hence, there must be a constant $C > 0$ such that

$$W_\phi(w, b) \leqslant C \tag{66}$$

for all $(w, b) \in \Omega$. This implies that if $\mu \in M_{\phi,f}^\Omega$ for $f \in V_\phi^\Omega$, then

$$\|f\|_{\mathcal{B}_\phi^\Omega} \leqslant \|\mu\|_{W_\phi,\Omega} \leqslant C \int_\Omega d|\mu|(w, b). \tag{67}$$

Taking the infimum over $\mu \in M_{\phi,f}^\Omega$ gives

$$\|f\|_{\mathcal{B}_\phi^\Omega} \leqslant C\|f\|_{V_\phi^\Omega}. \tag{68}$$

The combination of eq. (65) and eq. (67) shows the second statement. *Q.E.D.*

Note that proposition 3.1 excludes the higher order ReLU. Since the higher order ReLU satisfies

$$c^{\alpha}\sigma_{\alpha}(z) = \sigma_{\alpha}(cz) \tag{69}$$

for all $c > 0$ and $z \in \mathbb{R}$, we can restrict the size of the elements in Ω and compensate by increasing the measure. This allows us to write down a stronger version of proposition 3.1 for the higher order ReLU.

Proposition 3.2. *Let $\alpha \in \mathbb{N}$. If $\phi = \sigma_{\alpha}$, then*

$$\mathcal{B}^{\Omega}_{\sigma_{\alpha}} \simeq \mathcal{B}^{\mathbb{S}^{d+1}}_{\sigma_{\alpha}} = V^{\mathbb{S}^{d+1}}_{\sigma_{\alpha}}. \tag{70}$$

If Ω is a compact set, then we also have that

$$V^{\mathbb{S}^{d+1}}_{\sigma_{\alpha}} \cong V^{\Omega}_{\sigma_{\alpha}}. \tag{71}$$

Proof. Observe that $\mathcal{B}^{\mathbb{S}^{d+1}}_{\phi}$ and $V^{\mathbb{S}^{d+1}}_{\phi}$ have the same definition. This means that it is sufficient to show that both the Bach and Barron spaces for higher order ReLU can be expressed over $\mathbb{S}^{d+1}$ as well as over any other nonempty $\Omega \subseteq \mathbb{R}^{d+1}$.

We will first prove this for the Barron spaces, i.e prove eq. (70). If $\mu \in M^{\Omega}_{\phi,f}$ for $f \in \mathcal{B}^{\Omega}_{\sigma_{\alpha}}$, then the measure

$$d\gamma(w,b) = (\|w\| + |b|)^{\alpha} d\nu(w,b) \tag{72}$$

defined using the push forward

$$\nu = \Theta_{\#}(\mu) \tag{73}$$

along the map

$$\Theta : \Omega \to \mathbb{S}^{d+1}, \quad (w,b) \mapsto \left(\frac{w}{\|w\| + |b|}, \frac{b}{\|w\| + |b|}\right) \tag{74}$$

satisfies

$$\begin{aligned}
K^{\mathbb{S}^{d+1}}_{\sigma_{\alpha}}\gamma(x) &= \int_{\mathbb{S}^{d+1}} \sigma_{\alpha}(\langle x|w\rangle + b) d\gamma(w,b) \\
&= \int_{\mathbb{S}^{d+1}} \sigma_{\alpha}(\langle x|w\rangle + b)(\|w\| + |b|)^{\alpha} d\nu(w,b) \\
&= \int_{\Omega} \sigma_{\alpha}(\left\langle x \middle| \frac{w}{\|w\| + |b|} \right\rangle + \frac{b}{\|w\| + |b|})(\|w\| + |b|)^{\alpha} d\mu(w,b) \\
&= \int_{\Omega} \sigma_{\alpha}(\langle x|w\rangle + b) d\mu(w,b) \\
&= K^{\Omega}_{\sigma_{\alpha}}\mu(x) \\
&= f(x),
\end{aligned}$$

and thus

$$\|f\|_{\mathcal{B}^{\mathbb{S}^{d+1}}_{\sigma_{\alpha}}} \leqslant \|\gamma\|_{W_{\sigma_{\alpha}},\mathbb{S}^{d+1}} = \|\mu\|_{W_{\sigma_{\alpha}},\Omega}. \tag{75}$$

After taking the infimum over $\mu \in M_{\phi,f}$, we get

$$\|f\|_{\mathcal{B}^{\mathbb{S}^{d+1}}_{\sigma_{\alpha}}} \leqslant \|f\|_{\mathcal{B}^{\Omega}_{\sigma_{\alpha}}}. \tag{76}$$

It remains to show that

$$\|f\|_{\mathcal{B}^{\mathbb{S}^{d+1}}_{\sigma_\alpha}} \geqslant \|f\|_{\mathcal{B}^{\Omega}_{\sigma_\alpha}}. \tag{77}$$

This is immediate when $\mathbb{S}^{d+1} \subseteq \Omega$. When $\mathbb{S}^{d+1} \nsubseteq \Omega$, consider the biggest sphere $B \subseteq \mathbb{R}^{d+1}$ centered at the origin such that $B \subseteq \Omega$. Again, it follows immediately that

$$\|f\|_{\mathcal{B}^{B}_{\sigma_\alpha}} \geqslant \|f\|_{\mathcal{B}^{\Omega}_{\sigma_\alpha}}. \tag{78}$$

Hence, showing that

$$\|f\|_{\mathcal{B}^{\mathbb{S}^{d+1}}_{\sigma_\alpha}} \geqslant \|f\|_{\mathcal{B}^{B}_{\sigma_\alpha}} \tag{79}$$

is sufficient. The radius of $\mathbb{S}^{d+1}$ is one, and let r be the radius of B. Use this to define a new map

$$\Theta : B \to \mathbb{S}^{d+1},\ (w, b) \mapsto (w/r, b/r). \tag{80}$$

If $\mu \in M_{\sigma_\alpha, f}$ for $f \in \mathcal{B}^{\mathbb{S}^{d+1}}_{\sigma_\alpha}$, then the push forward

$$\nu(w, b) = \Theta_{\#}(r^\alpha \mu(w, b)) \tag{81}$$

satisfies

$$\begin{aligned} K^{B}_{\sigma_\alpha}\nu(x) &= \int_B \sigma_\alpha(\langle x|w\rangle + b)d\nu(w, b) \\ &= \int_{\mathbb{S}^{d+1}} \sigma_\alpha(\Big\langle x\Big|\frac{w}{r}\Big\rangle + \frac{b}{r})r^\alpha d\mu(w, b) \\ &= \int_{\mathbb{S}^{d+1}} \sigma_\alpha(\langle x|w\rangle + b)d\mu(w, b) \\ &= K^{\mathbb{S}^{d+1}}_{\sigma_\alpha}\mu(x) \\ &= f(x), \end{aligned}$$

and thus

$$\|f\|_{\mathcal{B}^{B}_{\sigma_\alpha}} \leqslant \|\nu\|_{W_{\sigma_\alpha}, B} = \|\mu\|_{W_{\sigma_\alpha}, \mathbb{S}^{d+1}}. \tag{82}$$

After taking the infimum over μ we get

$$\|f\|_{\mathcal{B}^{B}_{\sigma_\alpha}} \leqslant \|f\|_{\mathcal{B}^{\mathbb{S}^{d+1}}_{\sigma_\alpha}}. \tag{83}$$

Now, we will prove it for the Bach spaces, i.e. prove eq. (71). For this we have that Ω is compact. The strategy that we will use is similar to that for the Barron spaces.

Let R_Ω be the radius of the smallest closed ball $B_{R_\Omega}(0)$ centered at the origin such that $\Omega \subseteq B_{R_\Omega}(0)$. If $\mu \in M_{\sigma_\alpha, f}$ for $f \in V^{\Omega}_{\sigma_\alpha}$, then the measure

$$d\gamma(w, b) = (\|w\| + |b|)^\alpha d\nu(w, b) \tag{84}$$

defined using the push forward

$$\nu = \Theta_{\#}(\mu) \tag{85}$$

along the map

$$\Theta : \Omega \to \mathbb{S}^{d+1}, \quad (w, b) \mapsto (\frac{w}{\|w\| + |b|}, \frac{b}{\|w\| + |b|}) \tag{86}$$

satisfies

$$K_{\sigma_\alpha}^{\mathbb{S}^{d+1}}\gamma = K_{\sigma_\alpha}^{\Omega}\mu = f, \tag{87}$$

and thus

$$\|f\|_{V_{\sigma_\alpha}^{\mathbb{S}^{d+1}}} \leqslant \|\gamma\|_{rca(\mathbb{S}^{d+1})} \leqslant R_\Omega^\alpha \|\nu\|_{rca(\mathbb{S}^{d+1})} = R_\Omega^\alpha \|\mu\|_{rca(\Omega)}. \tag{88}$$

Taking the infimum over $\mu \in M_{\sigma_\alpha, f}$ gives

$$\|f\|_{V_{\sigma_\alpha}^{\mathbb{S}^{d+1}}} \leqslant R_\Omega^\alpha \|f\|_{V_{\sigma_\alpha}^{\Omega}}. \tag{89}$$

What remains is to show that

$$\|f\|_{V_{\sigma_\alpha}^{\mathbb{S}^{d+1}}} \geqslant C\|f\|_{V_{\sigma_\alpha}^{\Omega}} \tag{90}$$

for some constant $C > 0$. Let R_Ω be the radius of the biggest closed ball $B_{R_\Omega}(0)$ centered at the origin such that $B_{R_\Omega}(0) \subseteq \Omega$. Clearly

$$\|f\|_{V_{\sigma_\alpha}^{\Omega}} \leqslant \|f\|_{V_{\sigma_\alpha}^{B_{R_\Omega}(0)}}. \tag{91}$$

If $\mu \in M_{\sigma_\alpha, f}$ for $f \in V_{\sigma_\alpha}^{\mathbb{S}^{d+1}}$, then the push forward

$$\nu = \Theta_\#(R_\Omega^{-\alpha}\mu) \tag{92}$$

along the map

$$\Theta : \mathbb{S}^{d+1} \to B_{R_\Omega}(0),\ (w, b) \to (wR_\Omega, bR_\Omega) \tag{93}$$

satisfies

$$K_{\sigma_\alpha}^{B_{R_\Omega}(0)}\nu = K_{\sigma_\alpha}^{\mathbb{S}^{d+1}}\mu = f, \tag{94}$$

and thus

$$\|f\|_{V_{\sigma_\alpha}^{B_{R_\Omega}(0)}} \leqslant \|\nu\|_{rca(B_{R_\Omega}(0))} = R_\Omega^{-\alpha}\|\mu\|_{rca(\mathbb{S}^{d+1})}. \tag{95}$$

Taking the infimimum over $\mu \in M_{\sigma_\alpha, f}$ gives

$$\|f\|_{V_{\sigma_\alpha}^{B_{R_\Omega}(0)}} \leqslant R_\Omega^{-\alpha}\|f\|_{V_{\sigma_\alpha}^{\mathbb{S}^{d+1}}}. \tag{96}$$

The combination of eq. (91) and eq. (96) shows that eq. (90) holds. *Q.E.D.*

Note that in proposition 3.2 we purposefully demand that Ω is compact in order to have

$$V_{\sigma_\alpha}^{\mathbb{S}^{d+1}} \cong V_{\sigma_\alpha}^{\Omega}. \tag{97}$$

On unbounded Ω we see that

$$f(x) = \int_\Omega \sigma_\alpha(\langle x|w\rangle + b)d\mu(w, b) = \lim_{a\to\infty}\int_\Omega \sigma_\alpha(\langle x|wa\rangle + ba)d(a^{-\alpha}\mu)(w, b) \tag{98}$$

whilst

$$0 \leqslant \|f\|_{V_{\sigma_\alpha}^{\Omega}} \leqslant \lim_{a\to\infty}\left\|a^{-\alpha}\mu\right\|_{rca(\Omega)} = \lim_{a\to\infty}\left|a^{-\alpha}\right|\|\mu\|_{rca(\Omega)} = 0 \tag{99}$$

for all $\mu \in M_{\sigma_\alpha, f}$. This implies that on unbounded Ω we cannot find a constant $C > 0$ such that

$$\|f\|_{V_{\sigma_\alpha}^{\mathbb{S}^{d+1}}} \leqslant C\|f\|_{V_{\sigma_\alpha}^{\Omega}} \tag{100}$$

for all $f \in V_{\sigma_\alpha}^{\Omega}$. On bounded domains this construction also shows that the measures giving the smallest Bach norms will place the weights w and biases b on the boundary of the domain.

3.1 Completeness

The weight function W_ϕ not only makes the Barron spaces more realistic, it also makes them Banach spaces for arbitrary Ω. The Bach spaces, on the other hand, are only Banach spaces on compact Ω.

To prove that a space is Banach we need to show that it is a complete normed vector space. This is typically done by checking each of the axiomatic properties. In this case we know that

$$\|f\|_{V_\phi^\Omega} = \inf_{\mu \in M_{\phi,f}^\Omega} \|\mu\|_{rca(\Omega)} \tag{101}$$

for all $f \in V_\phi^\Omega$. At the same time we recall that the norm of the quotient space Z/Q, with Z a Banach space and Q are closed subspace thereof, is given by

$$\|[z]]\|_{Z/M} = \inf_{q \in [z]} \|q\|_Z \tag{102}$$

for each equivalent class $[z] \in Z/M$. This strongly suggests that V_ϕ^Ω can be identified with a quotient space of $rca(\Omega)$. We show in proposition 3.3 that this is indeed the case. Subsequently, we can use that completeness is a three-space-property to conclude that the Bach spaces must be Banach spaces. That completeness is a three-space-property is often seen a standard result, but we will include the proof for completeness. This can be found in lemma 3.0.1.

Lemma 3.0.1. *Let Z be a Banach space and Q a closed linear subspace such that Z/Q is a quotient space. If Z is complete, so is Z/Q.*

Proof. We will first show that Z is Banach if and only if for every sequence $(x_n)_{n\in\mathbb{N}}$ for which $\sum_{i=1}^\infty \|x_n\|_Z$ converges the sum $\sum_{i=1}^\infty x_n$ converges too.

Assume $\sum_{i=1}^\infty \|x_n\|_Z$ converges, then the partial sums $(\sum_{i=1}^k x_n)_{k\in\mathbb{N}}$ are Cauchy. Since Z is complete, they converge.

Conversely, suppose $(x_n)_{n\in\mathbb{N}}$ is Cauchy. Then there exists a relabelling $(x_{n_k})_{k\in\mathbb{N}}$ such that $\left\|x_{n_{k+1}} - x_{n_k}\right\|_Z < 2^{-k}$. This implies that the partial sums $(\sum_{i=1}^k \left\|x_{n_{i+1}} - x_{n_i}\right\|_Z)_{k\in\mathbb{N}}$ converge as $k \to \infty$. This in turn implies that the partial sums

$$\sum_{i=1}^k (x_{n_{i+1}} - x_{n_i}) = x_{n_{k+1}} - x_{n_1} \tag{103}$$

converge to some $x \in Z$. Hence, the relabelled sequence $(x_{n_k})_{k\in\mathbb{N}}$ converges to $x + x_{n_1}$. Since $(x_n)_{n\in\mathbb{N}}$ has a convergent subsequence $(x_{n_k})_{k\in\mathbb{N}}$, it must converge too. Z must be Banach, because $(x_n)_{n\in\mathbb{N}}$ was chosen arbitrarily.

Now what remains is to show that sequences $([x_n])_{n\in\mathbb{N}}$ with $[x_n] \in Z/Q$ and $\sum_{i=1}^\infty \|[x_n]\|_{Z/Q} < \infty$ indeed converge. For each of those sequences $([x_n])_{n\in\mathbb{N}}$ and each $n \in \mathbb{N}$ there exists a $y \in \mathbb{Q}$ such that

$$\|x_n - y_n\|_Z \leqslant \|[x_n]\|_{Z/Q} + \frac{1}{2^n}. \tag{104}$$

By construction $\sum_{n=1}^\infty \|x_n - y_n\|_Z < \infty$. This implies that the partial sums $(\sum_{n=1}^k x_n - y_n)_{k\in\mathbb{N}}$ form a Cauchy sequence. Since Z is complete, the partial sums $(\sum_{n=1}^k x_n - y_n)_{k\in\mathbb{N}}$ converge to some $z \in Z$ as $k \to \infty$.

Since $0 \in Q$, $\|[z]\|_{Z/Q} \leqslant \|z\|_Z$ for all $z \in Z$. Furthermore, since $y_n \in Q$, so is $\sum_{n=1}^k y_n$. Hence,

$$\begin{aligned}\left\|\sum_{n=1}^k [x_n] - [z]\right\|_{Z/Q} &= \left\|[\sum_{n=1}^k x_n - z]\right\|_{Z/Q} \\ &= \left\|[\sum_{n=1}^k x_n - z - \sum_{n=1}^k y_n]\right\|_{Z/Q} \\ &= \left\|[\sum_{n=1}^k (x_n - y_n) - z]\right\|_{Z/Q} \\ &\leqslant \left\|\sum_{n=1}^k (x_n - y_n) - z\right\|_Z \to 0\end{aligned}$$

as $k \to \infty$. This shows that the partial sums $(\sum_{n=1}^k [x_n]_{k\in\mathbb{N}}$ converge to $[z] \in Z/Q$ when $k \to \infty$.

Since the sequence $([x_n])_{n\in\mathbb{N}}$ was chosen arbitrarily, Z/M must be Banach. *Q.E.D.*

Now that we have shown that completeness carries over to the quotient space, we continue by proving that the Bach spaces are Banach.

Proposition 3.3. *If $\phi \in C(\mathbb{R})$, then $V_\phi^\Omega \simeq rca(\Omega)/M_{\phi,0}^\Omega$ using the isometric isomorphism*

$$T : V_\phi^\Omega \to rca(\Omega)/M_{\phi,0}^\Omega,\ f \mapsto Tf = M_{\phi,f}^\Omega, \tag{105}$$

and V_ϕ^Ω is a Banach space.

Proof. It is sufficient to show that

1. $M_{\phi,0}^\Omega$ is closed in $rca(\Omega)$,
2. the sets $M_{\phi,f}$ are sets generated from $M_{\phi,f}^\Omega$, i.e. for all $\mu \in M_{\phi,f}^\Omega$ it holds that $M_{\phi,f}^\Omega = [\mu] = \mu + M_{\phi,0}^\Omega$.

If these two hold, then

$$\|f\|_{V_\phi^\Omega} = \inf_{\mu\in M_{\phi,f}^\Omega} \|\mu\|_{1,\Omega} = \inf_{\mu\in M_{\phi,f}^\Omega} \|\mu\|_{rca(\Omega)} = \|[\mu]\|_{rca(\Omega)/M_{\phi,0}^\Omega} = \|Tf\|_{rca(\Omega)/M_{\phi,0}^\Omega}. \tag{106}$$

This implies that $\|T\| = 1$, which implies that V_ϕ^Ω is isometrically isomorphic to $rca(\Omega)/M_{\phi,0}^\Omega$. It then follows from lemma 3.0.1 that V_ϕ^Ω is a Banach space.

To show 1., observe that K_ϕ^Ω is linear and that $M_{\phi,0}^\Omega$ is the kernel of K_ϕ^Ω. As a consequence, $M_{\phi,0}^\Omega$ is a linear subspace of $rca(\Omega)$. Consider now a sequence of $\mu_n \in M_{\phi,0}$ such that

$$\mu_n \xrightarrow{\|\cdot\|_{rca(\Omega)}} \mu \tag{107}$$

for some $\mu \in rca(\Omega)$. Then for all $x \in \mathcal{X}$

$$\lim_{n\to\infty} |K_\phi \mu_n(x) - K_\phi \mu(x)| = \lim_{n\to\infty} |K_\phi(\mu_n - \mu)(x)| \leqslant \|\phi\|_{C(\mathbb{R})} \lim_{n\to\infty} \|\mu_n - \mu\|_{rca(\Omega)} = 0. \tag{108}$$

Since $K_\phi \mu_n(x) = 0$ for each $x \in \mathcal{X}$ and $n \in \mathbb{N}$, this upper bound implies that $|K_\phi \mu(x)| = 0$. Hence, $\mu \in M^\Omega_{\phi,0}$, which implies that $M^\Omega_{\phi,0}$ is closed.

To show 2., consider an arbitrary $\mu \in M^\Omega_{\phi,f}$ for an $f \in V^\Omega_\phi$. If $\nu \in [\mu] := \mu + M^\Omega_{\phi,0}$, then $\nu = \mu + \rho$ for some $\rho \in M^\Omega_{\phi,0}$ and

$$K_\phi \nu = K_\phi \mu + K_\phi \rho = f + 0 = f. \tag{109}$$

Hence, $\nu \in M^\Omega_{\phi,f}$. Conversely, if $\nu \in M^\Omega_{\phi,f}$, then

$$K_\phi(\nu - \mu) = K_\phi \nu - K_\phi \mu = f - f = 0. \tag{110}$$

Hence, $\nu - \mu \in M^\Omega_{\phi,0}$, which implies that $\nu \in [\mu] = \mu + M^\Omega_{\phi,0}$. *Q.E.D.*

We can prove that the Barron spaces are Banach spaces using the fact that the Bach spaces are.

Proposition 3.4. *If ϕ is an activation function for which W_ϕ is defined, then $\mathcal{B}^\Omega_\phi$ is Banach.*

Proof. If Ω is a compact set, then it follows from proposition 3.2 and proposition 3.1 that $\mathcal{B}^\Omega_\phi \cong V^\Omega_\phi$. We have just shown in proposition 3.3 that V^Ω_ϕ is a Banach space. Hence, $\mathcal{B}^\Omega_\phi$ is Banach too.

If Ω is not a compact set, we use a different strategy to prove completeness. We prove that this strategy works for $\phi \in C^{0,1}(\mathbb{R})$. The remaining cases can be done similarly.

Consider the space

$$r\tilde{c}a(\Omega) = \left\{ \mu \in rca(\Omega) \;\middle|\; \int_\Omega 1 + \|w\| + |b| d|\mu|(a,b) < \infty \right\}. \tag{111}$$

That $r\tilde{c}a(\Omega)$ is a vector space follows directly from that $rca(\Omega)$ is a vector space. Observe that for each $\mu \in M^\Omega_{\phi,f}$ for $f \in \mathcal{B}^\Omega_\phi$ we have that $\mu \in r\tilde{c}a(\Omega)$. Following the same arguments as in proposition 3.3 but replacing $rca(\Omega)$ with $r\tilde{c}a(\Omega)$ and V^Ω_ϕ with $\mathcal{B}^\Omega_\phi$, tells us that

$$\mathcal{B}^\Omega_\phi \simeq r\tilde{c}a(\Omega)/M^\Omega_{\phi,0}. \tag{112}$$

So to show that $\mathcal{B}^\Omega_\phi$ is Banach, it is sufficient to show that $r\tilde{c}a(\Omega)$ is complete.

Consider a Cauchy sequence $(\mu_n)_{n\in\mathbb{N}}$ with $\mu_n \in r\tilde{c}a(\Omega)$. $(\mu_n)_{n\in\mathbb{N}}$ form a Cauchy sequence if and only if the measures ν_n given by

$$d\nu_n(w,b) = (1 + \|w\| + |b|) d\mu_n(w,b) \tag{113}$$

form a Cauchy sequence in $rca(\Omega)$. Since $rca(\Omega)$ is complete, there must be a $\nu \in rca(\Omega)$ such that

$$\nu_n \xrightarrow{\|\cdot\|_{rca(\Omega)}} \nu. \tag{114}$$

Since $(1 + \|w\| + |b|)$ is invertible, finite and well-defined for all $(w, b) \in \Omega$, it must hold that $(\mu_n)_{n\in\mathbb{N}}$ (strongly) converges to $\mu \in r\check{c}a(\Omega)$ given by

$$d\mu(w, b) = \frac{1}{1 + \|w\| + |b|} d\nu(w, b). \tag{115}$$

$r\check{c}a(\Omega)$ must be complete, because the sequence (μ_n) was arbitrary. *Q.E.D.*

3.2 More general weight functions

The weight function W_ϕ used in the definition of the Barron norm penalises the measures linearly. In some cases we might want to weigh the size of the measure in a different way than linearly. We could adapt the definition of $\|\cdot\|_{W,\Omega}$ to

$$\|\mu\|_{W,\Omega,p} := \int_\Omega W(w, b) d|\mu|^p(w, b) \tag{116}$$

for $\mu \in rca(\Omega)$. However, this still limits the weight functions we can use. A more general way is to go from measures to probability measures. Since all measures $\mu \in rca(\Omega)$ have finite total variation, we can divide a measure $\mu \in rca(\Omega)$ by its total variation $\|\mu\|_{rca(\Omega)}$ to get a probability measure $\pi \in \mathbb{P}(\Omega)$. Clearly, the measures μ and π are zero on the same sets. Hence, the Radon-Nikodym derivative $\frac{d\mu}{d\pi}$ must exist. This allows us to split the measure μ into a function and a probability measure. To make this more precise consider the map

$$T_{\mathbb{P}} : rca(\Omega) \to \Big((\Omega \to \mathbb{R}) \times \mathbb{P}(\Omega)\Big),\ \mu \mapsto \left(\frac{d\mu}{d(\frac{|\mu|}{\|\mu\|_{rca(\Omega)}})}, \frac{|\mu|}{\|\mu\|_{rca(\Omega)}}\right). \tag{117}$$

$T_{\mathbb{P}}\mu$ is well-defined for all $\mu \in rca(\Omega)$, and $(a, \pi) := T_{\mathbb{P}}$ satisfy

$$d\mu(w, b) = a(w, b) d\pi(w, b) \tag{118}$$

for all $(w, b) \in \Omega$. Furthermore, the total variation measure $|\mu|$ satisfies

$$d|\mu|(w, b) = |a(w, b)| d\pi(w, b). \tag{119}$$

In eq. (118) π is a probability measure, and a can be seen as an associated density. With this formulation we can write

$$\|\pi\|_{\tilde{W},\Omega} = \int_\Omega \tilde{W}(a(w, b), w, b) d\pi(w, b). \tag{120}$$

Since the term $a(w, b)$ corresponds to the size of the measure, we can effectively penalise it by choosing $\tilde{W}$. To work with the infinitely wide neural network spaces that use this formulation, we define the tilde versions of K^Ω_ϕ, $M^\Omega_{\phi,f}$ and $\mathcal{N}^\Omega_{\phi,W}$.

***Definition* 5.** *Let ϕ be a fixed activation function, $\tilde{W} : \mathbb{R} \times \Omega \to \mathbb{R}$ be non-negative and set*

$$\tilde{K}^\Omega_\phi : \Big((\Omega \to \mathbb{R}) \times rca(\Omega)\Big) \to L^2(\mathcal{X}, \rho),\ (a, \pi) \mapsto \left(x \mapsto \int_\Omega a(w, b)\phi(\langle x|w\rangle + b) d\pi(w, b)\right)$$

$$\tilde{M}^\Omega_{\phi,f} = \left\{(a, \pi) = T_{\mathbb{P}}\mu \;\middle|\; \mu \in rca(\Omega),\ \forall x \in \mathcal{X} :\ f(x) = \tilde{K}^\Omega_\phi T_{\mathbb{P}}\mu(x)\right\}.$$

The special case that $f(x) = 0$ is denoted $\tilde{M}^{\Omega}_{\phi,0}$. The space $(L^{\Omega,\mathcal{X}}_{\phi,W,p}, \|\cdot\|_{L^{\Omega,\mathcal{X}}_{\phi,W,p}})$ given by

$$\tilde{\mathcal{N}}^{\Omega}_{\phi,W} = \left\{ f : \mathcal{X} \to \mathbb{R} \;\middle|\; \|f\|_{\tilde{\mathcal{N}}^{\Omega}_{\phi,W}} < \infty \right\}$$
$$\|f\|_{\tilde{\mathcal{N}}^{\Omega}_{\phi,W}} = \inf_{(a,\pi)\in\tilde{M}^{\Omega}_{\phi,f}} \int_{\Omega} \tilde{W}(a(w,b),w,b)d\pi(w,b)$$

is called a tilde infinitely wide neural network space. If $\tilde{\mathcal{N}}^{\Omega}_{\phi,1}$ is written, it is meant that $W(\omega) = 1$.

Remark. *In the definition of the tilde spaces we have used $(a,\pi) = T_{\mathbb{P}}\mu$. Although one could argue that one should consider all densities $a : \Omega \to \mathbb{R}$ from some L^p space with π as measure, this is not needed. Since the proof for this equivalence is similar to many proofs in this work and the notation for this all density function version is more cumbersome to write, it will not be used.*

If $\tilde{W}$ is absolutely homogeneous in its first argument, the tilde versions match the originals.

Proposition 3.5. *Let ϕ be a fixed activation function. If $\tilde{W} : \mathbb{R} \times \Omega \to \mathbb{R}$ a non-negative function which can be written as*

$$\tilde{W}(c,w,b) = |c|W(w,b) \tag{121}$$

for some non-negative $W : \Omega \to \mathbb{R}$ for all $(c,(w,b)) \in \mathbb{R} \times \Omega$, then $\tilde{\mathcal{N}}^{\Omega}_{\phi,W} \simeq \mathcal{N}^{\Omega}_{\phi,W}$.

Proof. Let $\mu \in M^{\Omega}_{\phi,f}$ for $f \in \mathcal{N}^{\Omega}_{W,\phi}$, then the pair $(a,\pi) = T_{\mathbb{P}}\mu$ satisfies

$$\tilde{K}^{\Omega}_{\phi}(a,\pi) = K^{\Omega}_{\phi}\mu = f, \tag{122}$$

and thus

$$\|f\|_{\tilde{\mathcal{N}}^{\Omega}_{\phi,W}} \leqslant \int_{\Omega} \tilde{W}(a(w,b),w,b)d\pi(w,b) = \int_{\Omega} |a(w,b)|W(w,b)d\pi(w,b) = \int_{\Omega} W(w,b)d|\mu|(w,b). \tag{123}$$

Taking the infimum of $\mu \in M^{\Omega}_{\phi,f}$ gives

$$\|f\|_{\tilde{\mathcal{N}}^{\Omega}_{\phi,W}} \leqslant \|f\|_{\mathcal{N}^{\Omega}_{\phi,W}}. \tag{124}$$

Let $(a,\pi) \in \tilde{M}^{\Omega}_{\phi,f}$ for $f \in \tilde{\mathcal{N}}^{\Omega}_{\phi,W}$, then the measure μ given by

$$d\mu(w,b) = a(w,b)d\pi(w,b) \tag{125}$$

for all $(w,b) \in \Omega$ satisfies

$$K^{\Omega}_{\phi}\mu = \tilde{K}^{\Omega}_{\phi}(a,\pi) = f \tag{126}$$

and thus

$$\|f\|_{\mathcal{N}^{\Omega}_{\phi,W}} \leqslant \int_{\Omega} W(w,b)d\mu(w,b) = \int_{\Omega} |a(w,b)|W(w,b)d\pi(w,b) = \int_{\Omega} \tilde{W}(a(w,b),w,b)d\pi(w,b). \tag{127}$$

Taking the infimum of $(a,\pi) \in \tilde{M}^{\Omega}_{\phi,f}$ gives

$$\|f\|_{\mathcal{N}^{\Omega}_{\phi,W}} \leqslant \|f\|_{\tilde{\mathcal{N}}^{\Omega}_{\phi,W}}. \tag{128}$$

Q.E.D.

If $\tilde{W}$ is higher order absolutely homogeneous, then we get a nested structure.

Proposition 3.6. *Let ϕ be a fixed activation function. If $\tilde{W}_p : \mathbb{R} \times \Omega \to \mathbb{R}$ is a non-negative function which can be written as*

$$\tilde{W}_p(c, w, b) = |c|^p W(w, b) \tag{129}$$

for some non-negative $W \in L^\infty(\Omega)$ for all $(c, (w, b)) \in \mathbb{R} \times \Omega$, then $\tilde{\mathcal{N}}^\Omega_{\phi,\tilde{W}_q} \subseteq \tilde{\mathcal{N}}^\Omega_{\phi,\tilde{W}_p}$ for all $q \geqslant p \geqslant 1$.

Proof. Suppose $f \in \tilde{\mathcal{N}}^\Omega_{\phi,W_q}$, then there must be a pair $(a, \pi) \in \tilde{M}^\Omega_{\phi,f}$ such that $f = \tilde{K}^\Omega_\phi(a, \pi)$. For this pair (a, π) we have that

$$\begin{aligned}
\|a\|^p_{L^p(\Omega,\pi,W)} &= \int_\Omega |a(w,b)|^p W(w,b) d\pi(w,b) \\
&= \int_{\Omega, |a(w,b)| \leqslant 1} |a(w,b)|^p W(w,b) d\pi(w,b) + \int_{\Omega, |a(w,b)|>1} |a(w,b)|^p W(w,b) d\pi(w,b) \\
&\leqslant \int_{\Omega, |a(w,b)| \leqslant 1} W(w,b) d\pi(w,b) + \int_{\Omega, |a(w,b)|>1} |a(w,b)|^q W(w,b) d\pi(w,b) \\
&\leqslant \|W\|_{L^\infty(\Omega)} + \|a\|^q_{L^q(\Omega,\pi,W)} < \infty.
\end{aligned}$$

Hence, $(a, \pi) \in \tilde{M}^\Omega_{\phi,f}$, and thus $f \in \tilde{\mathcal{N}}^\Omega_{\phi,W_p}$. *Q.E.D.*

Note that in proposition 3.6 we have $\tilde{\mathcal{N}}^\Omega_{\phi,\tilde{W}_q} \subseteq \tilde{\mathcal{N}}^\Omega_{\phi,\tilde{W}_p}$ and not $\tilde{\mathcal{N}}^\Omega_{\phi,\tilde{W}_q} \hookrightarrow \tilde{\mathcal{N}}^\Omega_{\phi,\tilde{W}_p}$. This is because there is no single constant $C > 0$ such that

$$\|f\|_{\tilde{\mathcal{N}}^\Omega_{\phi,\tilde{W}_p}} \leqslant C \|f\|_{\tilde{\mathcal{N}}^\Omega_{\phi,\tilde{W}_q}} \tag{130}$$

for all $\tilde{\mathcal{N}}^\Omega_{\phi,\tilde{W}_q}$. If take W to be a constant $C > 0$ in proposition 3.6, then we know from Hölder's inequality that

$$\|a\|_{L^p(\Omega,\pi)} \leqslant \|a\|_{L^q(\Omega,\pi)} \tag{131}$$

for all $(a, \pi) \in M^\Omega(\phi, f)$ for $f \in \tilde{\mathcal{N}}^\Omega_{\phi,\tilde{W}_q}$ and $1 \leqslant p \leqslant q$. However,

$$\|f\|_{\tilde{\mathcal{N}}^\Omega_{\phi,\tilde{W}_p}} \leqslant C \|a\|^p_{L^p(\Omega,\pi)}. \tag{132}$$

The combination of eq. (131) and eq. (132) give the inequality

$$\|f\|_{\tilde{\mathcal{N}}^\Omega_{\phi,\tilde{W}_p}} \leqslant \|f\|^{q/p}_{\tilde{\mathcal{N}}^\Omega_{\phi,\tilde{W}_q}} \tag{133}$$

after taking the infimum over $(a, \pi) \in M^\Omega(\phi, f)$. This shows that even in this simple case we do not have an embedding. If we allow for p'th roots around the intergral in the norm, we see that an embedding is possible. This case has been considered in [E et al., 2021] for the case that $\Omega = \mathbb{S}^{d+1}$ and $\phi = \sigma_1$. They show that when we consider using the norm

$$\inf_{(a,\pi) \in \tilde{M}^\Omega_{\phi,f}} \|a\|_{L^p(\Omega,\pi)} \tag{134}$$

instead of

$$\inf_{(a,\pi)\in\tilde{M}^{\Omega}_{\phi,f}} \|a\|^p_{L^p(\Omega,\pi)} \tag{135}$$

that each $p \geqslant 1$ gives the same value for the norm. This result is very specific to absolutely homogeneous activation functions. This includes all the higher order ReLU. Hence, for those functions we can use the $p = 1$, $p = 2$ or $p = \infty$ cases, which are typically the most simple cases.

Another result for the Barron spaces with the ReLU activation function is that they have a kind of ℓ^2 norm.

Proposition 3.7. *If $\phi = \sigma_1$, $\Omega = \mathbb{R}^d$ and*

$$\tilde{W} : \mathbb{R} \times \Omega \to \mathbb{R},\ (a, w, b) \mapsto |a|^2 + (\|w\| + |b|)^2, \tag{136}$$

then $\tilde{\mathbb{N}}^{\Omega}_{\phi,\tilde{W}} \cong \mathcal{B}^{\Omega}_{\phi}$.

Proof. We will show that $\tilde{\mathbb{N}}^{\Omega}_{\phi,\tilde{W}} \cong \tilde{\mathbb{N}}^{\Omega}_{\phi,\tilde{W}_\phi}$ with

$$\tilde{W}_\phi(a, w, b) = |a| W_\phi(w, b). \tag{137}$$

The proof then follows from proposition 3.5.

Due to the homogeneity of σ_1 we have that

$$f(x) = \int_\Omega a(w,b)\phi(\langle x|w\rangle + b)d\pi(w,b) = \int_\Omega \frac{a(w,b)}{\gamma}\phi(\langle x|\gamma w\rangle + \gamma b)d\pi(w,b) \tag{138}$$

for all $\gamma > 0$ for all $f \in \tilde{\mathbb{N}}^{\Omega}_{\phi,\tilde{W}_\phi}$. If we consider the map

$$\Theta^\gamma : \Omega \to \Omega,\ (w, b) \mapsto (\gamma w, \gamma b), \tag{139}$$

then we have $(\frac{a}{\gamma}, \Theta^\gamma_\# \pi) \in \tilde{M}^{\Omega}_{\phi,f}$ whenever $(a, \pi) \in \tilde{M}^{\Omega}_{\phi,f}$. The value of γ which minimises the associated norm is the value that minimises the function

$$g(\gamma) = \left|\frac{a}{\gamma}\right|^2 + (\|\gamma w\| + |\gamma b|)^2 \tag{140}$$

for fixed $(a, w, b) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$. This is achieved for

$$\gamma = \sqrt{\frac{|a|}{\|w\| + |b|}}. \tag{141}$$

Therefore,

$$\|f\|_{\tilde{\mathbb{N}}^{\Omega}_{\phi,\tilde{W}}} \leqslant \int_\Omega \left|\frac{a(w,b)}{\gamma}\right|^2 + (\|w\| + |b|)^2 d\Theta^\gamma_\# \pi(w,b) = \int_\Omega |a(w,b)|(\|w\| + |b|)d\pi(w,b). \tag{142}$$

Taking the infimum over $(a, \pi) \in \tilde{M}^{\Omega}_{\phi,f}$ gives

$$\|f\|_{\tilde{\mathbb{N}}^{\Omega}_{\phi,\tilde{W}}} \leqslant \|f\|_{\tilde{\mathbb{N}}^{\Omega}_{\phi,\tilde{W}_\phi}}. \tag{143}$$

For the other way around we use a different map

$$\tilde{\Theta}^{\gamma} : \Omega \to \Omega,\ (w, b) \mapsto (\frac{w}{\gamma}, \frac{b}{\gamma}). \tag{144}$$

If $(a, \pi) \in \tilde{M}^{\Omega}_{\phi,f}$ for $f \in \tilde{\mathbb{N}}^{\Omega}_{\phi,\tilde{W}}$, then by similar arguments as before $(\gamma, \tilde{\Theta}^{\gamma}_{\#}\pi) \in \tilde{M}^{\Omega}_{\phi,f}$. Hence,

$$\|f\|_{\tilde{\mathbb{N}}^{\Omega}_{\phi,\tilde{W}_{\phi}}} \leqslant \int_{\Omega} |\gamma a(w, b)|(\|w\| + |b|)d\Theta^{\gamma}_{\#}\pi(w, b) = \int_{\Omega} |a(w, b)|^2 + (\|w\| + |b|)^2 d\pi(w, b). \tag{145}$$

Taking the infimum over $(a, \pi) \in \tilde{M}^{\Omega}_{\phi,f}$ gives

$$\|f\|_{\tilde{\mathbb{N}}^{\Omega}_{\phi,\tilde{W}_{\phi}}} \leqslant \|f\|_{\tilde{\mathbb{N}}^{\Omega}_{\phi,\tilde{W}}}. \tag{146}$$

Q.E.D.

This result only works when $\Omega = \mathbb{R}^d$. When this is not the case, we run into problems whenever the minimising probability measure π has an associated density a that satisfies

$$a(w, b) > \|w\| + |b| \tag{147}$$

for some $(w, b) \in \Omega$. The push forward maps Θ^{γ} are then not guaranteed to map back into Ω, which leads to invalid probability measures.

3.3 Reproducing kernel Hilbert spaces

The tilde formulation with a probability measure $\pi \in \mathbb{P}(\Omega)$ and a density $a : \Omega \to \mathbb{R}$ allows us to establish a link between the reproducing kernel Hilbert spaces.

***Definition* 6** (Reproducing kernel Hilbert spaces)**.** *Let A be some arbitrary set, and let*

$$k : A \times A \to \mathbb{R} \tag{148}$$

be a symmetric, positive definite function. k is called a kernel, and the Hilbert space $\mathcal{H}$ consisting of functions of the form

$$f : A \to \mathbb{R},\ x \mapsto \langle k(x, \cdot)|f\rangle_{\mathcal{H}} \tag{149}$$

is called a reproducing kernel Hilbert space, RKHS for short.

In order to show the link we will use the *kernel trick*.

Lemma 3.0.2. *[Kernel trick] If $u : A \to B$ for an arbitrary set A and a Hilbert space B, then*

$$k(x, y) = \langle u(x)|u(y)\rangle_B \tag{150}$$

is the kernel of a reproducing kernel Hilbert space.

With this lemma we can construct the reproducing kernel Hilbert spaces of interest.

Proposition 3.8. *For a fixed probability measure $\pi \in \mathbb{P}(\Omega)$, a nonzero measurable weight function W and an activation function $\phi \in L^2(\Omega, \pi)$, the function*

$$k(x,y) = \int_\Omega \phi(\langle x|w\rangle + b)\phi(\langle x|w\rangle + b)W^{-1}(w,b)d\pi(w,b) \tag{151}$$

is the kernel of a reproducing kernel Hilbert space.

Proof. If we consider a fixed probability measure $\pi \in \mathbb{P}(\Omega)$, then $L^2(\Omega, \pi, W)$ is a Hilbert Space and the function

$$u : \mathcal{X} \to L^2(\Omega, \pi, W),\ x \mapsto \Big((w,b) \mapsto \phi(\langle x|w\rangle + b)W^{-1}(w,b)\Big) \tag{152}$$

satisfies the requirements for lemma 3.0.2. This means

$$k(x,y) = \langle u(x)|u(y)\rangle_{L^2(\Omega,\pi,W)} = \int_\Omega \phi(\langle x|w\rangle + b)\phi(\langle y|w\rangle + b)W^{-1}(w,b)d\pi(w,b) \tag{153}$$

is the kernel of a reproducing kernel Hilbert space. *Q.E.D.*

We will refer to the map u of proposition 3.8 by $k^W_{\phi,\pi}$, and to the RKHS as $\mathcal{H}_{k^W_{\phi,\pi}}$. By definition, for each function f in the reproducing kernel Hilbert space $\mathcal{H}_{k^W_\pi}$, there is an $a \in L^2(\Omega, \pi, W)$ such that

$$f(x) = \left\langle a \middle| k^W_{\phi,\pi} \right\rangle_{L^2(\Omega,\phi,W)} = \int_\Omega a(w,b)\phi(\langle x|w\rangle + b)d\pi(w,b) \tag{154}$$

and the corresponding norm is

$$\|f\|_{\mathcal{H}_{k^W_{\phi,\pi}}} = \inf_{a \in L^2(\Omega,\pi,W)} \|a\|_{L^2(\Omega,\pi,W)} = \inf_{a \in L^2(\Omega,\pi,W)} \int_\Omega |a(w,b)|^2 W(w,b)d\pi(w,b). \tag{155}$$

At first glance this does not look like the norm of a $\tilde{\mathcal{N}}^\Omega_{\phi,\tilde{W}}$ due to the L^2 norm in the infimum, but if we take

$$\tilde{W}(a,w,b) = |a|^2 W(w,b), \tag{156}$$

then we see that

$$\|f\|_{\tilde{\mathcal{N}}^\Omega_{\phi,\tilde{W}}} = \inf_{\pi \in \mathbb{P}(\Omega)} \|f\|_{\mathcal{H}_{k^W_{\phi,\pi}}} \tag{157}$$

for all $f \in \tilde{\mathcal{N}}^\Omega_{\phi,\tilde{W}}$. This shows that $\tilde{\mathcal{N}}^\Omega_{\phi,\tilde{W}}$ is a union of RKHS, and that the norm of f in that space corresponds to the RKHS that can best represent f. Since the Bach and Barron spaces have

$$\tilde{W}(a,w,b) = |a|W(w,b) \tag{158}$$

with their respective weight functions W, they can be seen as L^1-like version of a union of RKHS. Note that this link is not a new result. A similar result has been shown in [E et al., 2021]. The difference between what was proven before and what is shown here is that we consider other weight functions than $W(w,b) = 1$.

3.4 Activation Functions

Up to this point we have mainly looked at the weight function. We will continue with looking at the activation function. In other works on Barron spaces the focus was on ReLU. This was done because ReLU has the nice property of homogeneity. Activation functions in general do not have this nice property. In this section we discuss relations between various spaces with different activation functions. We will start by perturbing the activation function a little, and work our way up to showing that the Barron space with ReLU as activation function can approximate the most functions of any activation function in use.

First, we show what happens when the activation function or its input is scaled or translated. We will do that by proving 4 lemmas in a row.

Lemma 3.0.3 (Consistency – scaling)**.** *If $\phi(x) = c\psi(x)$ for some activation function ψ and $c \in \mathbb{R}\backslash\{0\}$, then $\mathcal{N}_{\phi,W}$ is isomorphic to $\mathcal{N}_{\psi,W}$.*

Proof. Take $\mu \in M^{\Omega}_{\phi,f}$ for $f \in \mathcal{N}^{\Omega}_{\phi,W}$ and set $\nu = c\mu$, then

$$K^{\Omega}_{\psi}\nu = K^{\Omega}_{\phi}\mu = f. \tag{159}$$

Hence, $\nu \in M_{\psi,f}$. Furthermore,

$$\|f\|_{\mathcal{N}^{\Omega}_{\psi,W}} \leqslant \|\nu\|_{W,\Omega} = \int_{\Omega} W(w,b)d|\nu|(w,b) = |c| \int_{\Omega} W(w,b)d|\mu|(w,b) = |c|\|\mu\|_{W,\Omega}. \tag{160}$$

Taking the infimum over $\nu \in M_{\psi,f}$ gives

$$\|f\|_{\mathcal{N}^{\Omega}_{\psi,W}} \leqslant |c|\|f\|_{\mathcal{N}^{\Omega}_{\phi,W}}, \tag{161}$$

and thus $f \in \mathcal{N}_{\psi,W}$.

For the reverse direction observe that $\psi(x) = \frac{1}{c}\phi(x)$ and apply the just proven.

Q.E.D.

Lemma 3.0.4 (Consistency – magnification)**.** *Let $\phi(x) = \psi(cx)$ for some activation function ψ with $0 < c \leqslant 1$, and let W be a weight function. If W is sublinear, then $\mathcal{N}^{\Omega}_{\psi,W} \hookrightarrow \mathcal{N}^{\Omega}_{\phi,W}$. In particular, $\mathcal{N}^{\Omega}_{\psi,1} \hookrightarrow \mathcal{N}^{\Omega}_{\phi,1}$.*

Proof. Take $\mu \in M^{\Omega}_{\phi,f}$ for $f \in \mathcal{N}^{\Omega}_{\psi,W}$ and set $\nu = \Theta_{\#}\mu$ along the map

$$\Theta : \Omega \to c\Omega,\ (w,b) \mapsto (cw, cb), \tag{162}$$

then

$$K^{c\Omega}_{\phi}\nu = K^{\Omega}_{\psi}\mu = f. \tag{163}$$

Hence, $\nu \in M^{c\Omega}_{\phi,f}$. Combined with $c\Omega \subseteq \Omega$ we have $\nu \in M^{\Omega}_{\phi,f}$.

If W is sublinear, then

$$\begin{aligned}\|f\|_{\mathcal{N}_{\psi,W}} &\leqslant \int_{c\Omega} W(w,b)d|\nu|(w,b)\\ &= \int_{\Omega} W(cw,cb)d|\mu|(w,b)\\ &\leqslant \int_{\Omega} W(w,b)d|\mu|(w,b)\\ &= \|\mu\|_{W,\Omega}.\end{aligned} \tag{164}$$

Taking the infimum over $\mu \in M^{\Omega}_{\psi,f}$ gives

$$\|f\|_{\mathcal{N}^{\Omega}_{\phi,W}} \leqslant \|f\|_{\mathcal{N}^{\Omega}_{\psi,W}}, \tag{165}$$

and thus $f \in \mathcal{N}^{\Omega}_{\psi,W}$. *Q.E.D.*

Lemma 3.0.5 (Consistency – translation)**.** *Let $\phi(x) = \psi(x+c)$ for some activation function ψ with $c \in \mathbb{R}$, and set $\Omega_1 = U \times [-C, C]$ and $\Omega_2 = U \times [-C+c, C+c]$ with $U \subseteq \mathbb{R}^d$ and $C > 0$, then $V^{\Omega_1}_{\phi} \simeq V^{\Omega_2}_{\psi}$.*

Proof. Take $\mu \in M_{\phi,f}$ for $f \in V_{\phi}$ and set $\nu = \Theta_{\#}\mu$ along the map

$$\Theta : \Omega_1 \to \Omega_2,\ (w,b) \mapsto (w, b+c), \tag{166}$$

then

$$K^{\Omega_2}_{\psi}\nu = K^{\Omega_1}_{\phi}\mu = f. \tag{167}$$

Hence, $\nu \in M_{\psi,f}$. Furthermore,

$$\|f\|_{V^{\Omega_1}_{\psi}} \leqslant \|\nu\|_{rca(\Omega_1)} = \int_{\Omega_1} d|\nu|(w,b+c) = \int_{\Omega_2} d|\mu|(w,b) = \|\mu\|_{rca(\Omega_2)}. \tag{168}$$

Taking the infimum over $\mu \in M_{\phi,f}$ gives

$$\|f\|_{V^{\Omega_1}_{\psi}} \leqslant \|f\|_{V^{\Omega_2}_{\phi}}, \tag{169}$$

and thus $f \in V^{\Omega_1}_{\psi}$.

For the reverse embedding observe that $\psi(x) = \phi(x-c)$ and apply the just proven. *Q.E.D.*

Lemma 3.0.6 (Consistency – offset)**.** *Let ϕ be an activation function such that for some $(0,a) \in \Omega$ we have $\phi(a) \neq 0$, set $\psi(x) = c + \phi(x)$ with $c \in \mathbb{R}$ and let W be a weight function that is positive except for possibly at $(0,a)$, then $\mathcal{N}_{\psi,W}$ is isomorphic to $\mathcal{N}_{\phi,W}$. If W vanishes at $(0,a)$, then the isomorphism is isometric.*

Proof. Take $\mu \in M_{\phi,f}$ for $f \in \mathcal{N}^{\Omega}_{\phi,W}$ and set

$$d\nu(w,b) = c\frac{\mu(\Omega)}{\psi(a)}d\delta_{0,a}(w,b), \tag{170}$$

then

$$\begin{aligned}
f(x) &= \int_\Omega \phi(\langle x|w\rangle + b) d\mu(w,b) \\
&= \int_\Omega \psi(\langle x|w\rangle + b) + c d\mu(w,b) \\
&= \int_\Omega \psi(\langle x|w\rangle + b) d\mu + c\int_\Omega d\mu(w,b) \\
&= K_\psi \mu(x) + \frac{c\mu(\Omega)}{\psi(a)}\psi(a) \\
&= K_\psi \mu(x) + \frac{c\mu(\Omega)}{\psi(a)}\int_\Omega \psi(\langle x|w\rangle + b) d\delta_{0,a}(w,b) \\
&= K_\psi \mu(x) + K_\psi \nu(x) \\
&= K_\psi(\mu+\nu)(x).
\end{aligned}$$

Hence, $\mu + \nu \in M_{\psi,f}$. Furthermore,

$$\|\mu\|_{rca(\Omega)} \leqslant \frac{1}{\tilde{c}}\|\mu\|_{W,\Omega} \tag{171}$$

where $\tilde{c} > 0$ is the lower bound of W. This implies that

$$\begin{aligned}
\|f\|_{\mathcal{N}_{\psi,W}} &\leqslant \int_\Omega W(w,b) d|\mu+\nu|(w,b) \\
&\leqslant \int_\Omega W(w,b) d|\mu|(w,b) + \int_\Omega W(w,b) d|\nu|(w,b) \\
&= \int_\Omega W(w,b) d|\mu|(w,b) + \int_\Omega W(w,b) d\left|c\frac{\mu(\Omega)}{\psi(a)}\delta_{0,a}\right|(w,b) \\
&= \int_\Omega W(w,b) d|\mu|(w,b) + \left|\frac{c\mu(\Omega)}{\psi(a)}\right| \int_\Omega W(w,b) d\delta_{0,a}(w,b) \\
&= \int_\Omega W(w,b) d|\mu|(w,b) + \left|\frac{c}{\psi(a)}\right| W(0,a)|\mu(\Omega)| \\
&\leqslant \int_\Omega W(w,b) d|\mu|(w,b) + \left|\frac{c}{\psi(a)}\right| W(0,a)\|\mu\|_{rca(\Omega)} \\
&\leqslant \left(1 + \left|\frac{c}{\tilde{c}\psi(a)}\right| W(0,a)\right)\|\mu\|_{W,\Omega}.
\end{aligned}$$

Taking the infimum over $\mu \in M_{\phi,f}$ gives

$$\|f\|_{\mathcal{N}^\Omega_{\psi,W}} \leqslant \left(1 + \left|\frac{c}{\tilde{c}\psi(a)}\right| W(0,a)\right)\|f\|_{\mathcal{N}^\Omega_{\phi,W}}, \tag{172}$$

and thus $f \in \mathcal{N}^\Omega_{\psi,W}$.

To show that $\mathcal{N}^\Omega_{\psi,W} \hookrightarrow \mathcal{N}^\Omega_{\phi,W}$, it is sufficient to observe that $\phi(x) = -c + \psi(x)$ and to apply the just proven. *Q.E.D.*

Lemmas 3.0.3 till 3.0.6 show that basic operations on the activation functions do one of two things. Offsetting and scaling doesn't change what functions are in the spaces, but changes the norms with a certain factor. Magnification and translation change the underlying parameter space, but do not change the norms. These agree with the intuition of what would happen.

Secondly, we show what happens when an activation function can be constructed using the continuous version of a span of another activation function.

Lemma 3.0.7. *If there exists a $\gamma \in rca(U)$ with $U \subseteq \mathbb{R}^2$ such that $\phi(x) = \int_U \psi(\omega x + \beta) d\gamma(\omega, \beta)$, then $V_\phi^{\Omega_1} \hookrightarrow V_\psi^{\Omega_2}$ with*

$$\Omega_2 = \left\{ (\omega w, \omega b + \beta) \;\middle|\; (w, b) \in \Omega_1, (\omega, \beta) \in U \right\}. \tag{173}$$

Proof. Let $\mu \in M_{\phi,f}$ for $f \in B_\phi$. This means that

$$\begin{aligned} f(x) &= \int_{\Omega_1} \phi(\langle x|w\rangle + b) d\mu(w, b) \\ &= \int_{\Omega_1} \int_{\mathbb{R}^2} \psi(\omega(\langle x|w\rangle + b) + \beta) d\gamma(\omega, \beta) d\mu(w, b) \\ &= \int_{\Omega_1} \int_{\mathbb{R}^2} \psi(\langle x|\omega w\rangle + \omega b + \beta) d\gamma(\omega, \beta) d\mu(w, b) \\ &= \int_{\Omega_1 \times \mathbb{R}^2} \psi(\langle x|\omega w\rangle + \omega b + \beta) d(\gamma \otimes \mu)((w, b), (\omega, \beta)) \\ &= \int_{\Omega_2} \psi(\langle x|w\rangle + b) d\nu(w, b) \end{aligned}$$

where $\nu = \Theta_\#(\gamma \otimes \mu)$ is the pushforward along the map

$$\Theta : \Omega_1 \times \mathbb{R}^2 \to \Omega_2,\ ((w, b), (\omega, \beta)) \mapsto (\omega w, \omega b + \beta). \tag{174}$$

Hence,

$$\|f\|_{V_\psi} \leqslant \|\nu\|_{rca(\Omega)} \leqslant \|\gamma\|_{rca(\mathbb{R}^2)} \|\mu\|_{rca(\Omega)}. \tag{175}$$

Taking the infimum over $\mu \in M_{\phi,f}$ gives

$$\|f\|_{V_\psi} \leqslant \|\gamma\|_{rca(\mathbb{R}^2)} \|f\|_{V_\phi}. \tag{176}$$

Q.E.D.

Thirdly, we show that the Barron spaces with any sufficiently smooth activation function embed into Barron space for ReLU. The smoothness condition includes most of the activation functions used in practice.

Theorem 3.1. *If $\phi \in C^2(\mathbb{R})$, then $\mathcal{B}_\phi^\Omega \hookrightarrow \mathcal{B}_\sigma^{\mathbb{S}^{d+1}}$.*

Proof. We will show that his is true by showing that there exists an Ω_2 such that

$$\mathcal{B}_\phi^\Omega \hookrightarrow \mathcal{B}_\sigma^{\Omega_2}, \tag{177}$$

and then we can use proposition 3.2 to finish the proof.

To prove eq. (177) recall that $\mathcal{X}$ is compact and thus bounded. This means that there exists a constant $C_R > 0$ such that

$$|\langle x|w\rangle + b| \leqslant C_R(1 + \|w\| + |b|) \qquad \forall (x,(w,b)) \in \mathcal{X} \times \Omega_1. \tag{178}$$

Since $\phi \in \mathbb{C}^2(\mathbb{R})$, we have by theorem 4.1 that for given $(w,b) \in \Omega$

$$\phi(\langle x|w\rangle+b) = \phi(0)+\partial^1\phi(0)(\langle x|w\rangle+b)+\int_0^{C_R(1+\|w\|+|b|)} \partial^2\phi(t)\Big(\sigma(\langle x|w\rangle+b-t)-\sigma(\langle x|-w\rangle-b+t)\Big)dt \tag{179}$$

for all $x \in X$. After the change of coordinate $\theta_{w,b}u = t$ with $\theta_{w,b} = (1 + \|w\| + |b|)$, this becomes

$$\phi(\langle x|w\rangle+b) = \phi(0)+\partial^1\phi(0)(\langle x|w\rangle+b)+\int_0^{C_R} \frac{\partial^2\phi(\theta_{w,b}u)}{\theta_{w,b}}\Big(\sigma(\langle x|w\rangle+b-\theta_{w,b}u)-\sigma(\langle x|-w\rangle-b+\theta_{w,b}u)\Big)du. \tag{180}$$

Using lemma 3.0.6 we can take $\phi(0) = 0$ without loss of generality. Next, define

$$\Omega_2 = \left\{(\chi_1 w, \chi_2(b - \theta_{w,b}t)) \;\middle|\; (w,b) \in \Omega_1, t \in [0, C_R], \chi_i \in \{-1,1\}\right\}, \tag{181}$$

and observe that $\Omega_1 \subseteq \Omega_2$. Let $\mu \in M_{\phi,f}$ for $f \in \mathcal{B}_\phi^{\Omega_1}$, and consider the maps

$$\begin{aligned}
&\Theta^1 : \Omega_1 \to \Omega_2,\ (w,b) \mapsto (-w,-b)\\
&\Theta^2 : \Omega_1 \times [0,C]) \to \Omega_2,\ ((w,b),u) \mapsto (w, b - \theta_{w,b}u)\\
&\Theta^3 : \Omega_1 \times [0,C] \to \Omega_2,\ ((w,b),u) \mapsto (-w, -b + \theta_{w,b}u).
\end{aligned} \tag{182}$$

Use these maps to construct the measures

$$\begin{aligned}
\nu_1 &= \partial\phi(0)\mu\\
\nu_2 &= -\partial\phi(0)\Theta^1_\# \mu\\
d\nu_3((w,b),u) &= \frac{\partial^2\phi(\theta_{w,b}u)}{\theta_{w,b}} d(\mu \otimes \lambda)((w,b),u)\\
\nu_4 &= \Theta^2_\# \nu_3\\
\nu_5 &= -\Theta^3_\# \nu_3,
\end{aligned} \tag{183}$$

where λ is the Lebesgue measure on $[0,C]$. For the first two measures, ν_1 and ν_2, we have

$$\int_\Omega \sigma(\langle x|w\rangle + b) d(\nu_1 + \nu_2)(w,b) = \int_\Omega \partial\phi(0)(\langle x|w\rangle + b) d\mu(w,b). \tag{184}$$

For the last two measures, ν_4 and ν_5, we have

$$\int_\Omega \sigma(\langle x|w\rangle + b) d\nu_4(w,b) = \int_\Omega \int_0^{C_R} \frac{\partial^2\phi(\theta_{w,b}u)}{\theta_{w,b}} \sigma(\langle x|w\rangle + b - \theta_{w,b}u) du d\mu(w,b) \tag{185}$$

and

$$\int_\Omega \sigma(\langle x|w\rangle + b) d\nu_5(w,b) = -\int_\Omega \int_0^{C_R} \frac{\partial^2\phi(\theta_{w,b}u)}{\theta_{w,b}} \sigma(\langle x|-w\rangle - b + \theta_{w,b}u) du d\mu(w,b). \tag{186}$$

Hence, the measure

$$\nu = \nu_1 + \nu_2 + \nu_4 + \nu_5 \tag{187}$$

satisfies

$$f(x) = \int_\Omega \phi(\langle x|w\rangle + b)d\mu(w,b) = \int_\Omega \sigma(\langle x|w\rangle + b)d\nu(w,b). \tag{188}$$

Furthermore, for ν_1 and ν_2

$$\|\nu_1\|_{W_\sigma,\Omega_2} = \|\nu_2\|_{W_\sigma,\Omega_2} = \left|\partial^1\phi(0)\right|\|\mu\|_{W_\phi,\Omega_1} \leqslant \|\phi\|_{C^2(\mathbb{R})}\|\mu\|_{W_\phi,\Omega_1} \tag{189}$$

and for ν_4 and ν_5

$$\begin{aligned}
\|\nu_4\|_{W_\sigma,\Omega_2} = \|\nu_5\|_{W_\sigma,\Omega_2} &= \int_\Omega\int_0^{C_R}\left|\frac{\partial^2\phi(\theta_{w,b}u)}{\theta_{w,b}}\right|(\|w\| + |b - \theta_{w,b}u|)dud|\mu|(w,b) \\
&\leqslant \|\phi\|_{C^2(\mathbb{R})}\int_\Omega\int_0^{C_R}(\|w\| + |b - \theta_{w,b}u|)dud|\mu|(w,b) \\
&\leqslant \|\phi\|_{C^2(\mathbb{R})}\int_\Omega\int_0^{C_R}(\|w\| + |b| + |\theta_{w,b}u|)dud|\mu|(w,b) \\
&= \|\phi\|_{C^2(\mathbb{R})}\int_\Omega C_R(\|w\| + |b|) + \frac{1}{2}C_R^2|\theta_{w,b}|d|\mu|(w,b) \\
&\leqslant \max\{C_R, \frac{1}{2}C_R^2\}\|\phi\|_{C^2(\mathbb{R})}\|\mu\|_{W_\phi,\Omega_2}.
\end{aligned}$$

Combined this becomes

$$\|f\|_{\mathcal{B}_\sigma^{\Omega_2}} \leqslant \|\nu\|_{W_\sigma,\Omega_2} \leqslant \sum_{i=1,i\neq 3}^{5}\|\nu_i\|_{W_\sigma,\Omega_2} \leqslant 2\left(1 + \max\{C_R, \frac{1}{2}C_R^2\}\right)\|\phi\|_{C^2(\mathbb{R})}\|\mu\|_{W_\phi,\Omega_1}.s \tag{190}$$

Taking the infimum over $\mu \in M_{\phi,f}^\Omega$ gives

$$\|f\|_{\mathcal{B}_\sigma^{\Omega_2}} \leqslant 2\left(1 + \max\{C_R, \frac{1}{2}C_R^2\}\right)\|\phi\|_{C^2(\mathbb{R})}\|f\|_{\mathcal{B}_\phi^\Omega}. \tag{191}$$

This means that $f \in \mathcal{B}_\sigma^{\Omega_2}$, and that eq. (177) holds.

Proposition 3.2 implies that

$$\mathcal{B}_\sigma(\Omega_2) \hookrightarrow \mathcal{B}_\sigma(\mathbb{S}^{d+1}). \tag{192}$$

By transitivity of embeddings eq. (177) and eq. (192) show together that.

$$\mathcal{B}_\phi(\Omega) \hookrightarrow \mathcal{B}_\sigma(\mathbb{S}^{d+1}). \tag{193}$$

Q.E.D.

Lastly, we show that there exists an ordering of the Barron spaces with the higher-order ReLU activation functions. This result is distinct from theorem 3.1, since in general the higher order ReLU are smooth enough but not bounded.

Theorem 3.2. *If $\alpha \geqslant \beta \geqslant 1$ with $\alpha, \beta \in \mathbb{N}$, then $\mathcal{B}_{\sigma_\alpha}^\Omega \hookrightarrow \mathcal{B}_{\sigma_\beta}(\mathbb{S}^{d+1})$ for all $\Omega \subseteq \mathbb{R}^{d+1}$.*

Proof. Proposition 3.2 states that

$$\mathcal{B}^{\Omega}_{\sigma_\beta} \hookrightarrow \mathcal{B}^{\mathbb{S}^{d+1}}_{\sigma_\beta} \tag{194}$$

for all $\beta \in \mathbb{N}$ and $\Omega \subseteq \mathbb{R}^{d+1}$. If we can show that

$$\mathcal{B}^{\mathbb{S}^{d+1}}_{\sigma_{\beta+1}} \hookrightarrow \mathcal{B}^{\Omega}_{\sigma_\beta} \tag{195}$$

for some $\Omega \subseteq \mathbb{R}^{d+1}$, then

$$\mathcal{B}^{\Omega}_{\sigma_\alpha} \hookrightarrow \mathcal{B}^{\mathbb{S}^{d+1}}_{\sigma_\beta} \tag{196}$$

follows by alternatingly applying eq. (194) and eq. (195). Hence, to prove the statement it is sufficient to prove that eq. (195) holds.

By the assumptions on $\mathcal{X}$ there exists a closed ball $B_R(0)$ of radius $R > 0$ around the origin such that $\mathcal{X} \subset \overline{B_R}(0)$ for sufficiently large R. Set $C := 1 + R$, and

$$\Omega = \left\{ (w, b - t) \;\middle|\; (w, b) \in \mathbb{S}^{d+1}, t \in [0, C] \right\}. \tag{197}$$

Observe that

$$\sigma_{\beta+1}(y) = (\beta + 1) \int_0^C \sigma_\beta(y - t)dt \quad \forall y \in \mathbb{R} : |y| \leqslant C, \tag{198}$$

and that

$$|\langle x|w\rangle + b| \leqslant C \quad \forall x \in \mathcal{X} \tag{199}$$

for given $(w, b) \in \mathbb{S}^{d+1}$. Furthermore, recall that

$$\sigma_\beta(cx) = c^\beta \sigma_\beta(x) \quad \forall c \geqslant 0, x \in \mathbb{R} \tag{200}$$

by the homogeneity of σ. All this together means that for $\mu \in M^{\mathbb{S}^{d+1}}_{\sigma_{\beta+1},f}$ with $f \in \mathcal{B}^{\mathbb{S}^{d+1}}_{\sigma_{\beta+1}}$ and for all $x \in \mathcal{X}$ we have that

$$\begin{aligned}
f(x) &= \int_{\mathbb{S}^{d+1}} \sigma_{\beta+1}(\langle x|w\rangle + b)d\mu(w, b) \\
&= \int_{\mathbb{S}^{d+1}} (\beta + 1) \int_0^C \sigma_\beta(\langle x|w\rangle + b - t)dtd\mu(w, b) \\
&= \int_{\mathbb{S}^{d+1}} \int_0^C \sigma_\beta(\langle x|w\rangle + b - t)dtd((\beta + 1)\mu)(w, b) \\
&= \int_{[0,C]\times\mathbb{S}^{d+1}} \sigma_\beta(\langle x|w\rangle + b - t)d(\lambda \otimes (\beta + 1)\mu)(t, (w, b)) \\
&= \int_\Omega \sigma_\beta(\langle x|w\rangle + b)d\nu(w, b)
\end{aligned}$$

where λ is the Lebesque measure on $[0, C]$ and $\nu := \Theta_\#(\lambda \otimes (\beta + 1)\mu)$ is the pushforward of the product measure $\lambda \otimes (\beta + 1)\mu$ along the map

$$\Theta : \quad [0, C] \times \mathbb{S}^{d+1} \to \Omega, \quad (t, (w, b)) \mapsto (w, b - t). \tag{201}$$

Observe that

$$\|f\|_{\mathcal{B}^{\mathbb{S}^{d+1}}_{\sigma_{\beta+1}}} \leqslant \int_\Omega (\|w\|_1 + |b|)^\beta d\nu(w, b)$$

$$
\begin{aligned}
&= (\beta+1)\int_{\mathbb{S}^{d+1}}\int_0^C (\|w\|_1 + |b-t|)^\beta dt d|\mu|(w,b)\\
&\leqslant (\beta+1)\int_{\mathbb{S}^{d+1}}\int_0^C (\|w\|_1 + |b| + |t|)^\beta dt d|\mu|(w,b)\\
&= (\beta+1)\int_{\mathbb{S}^{d+1}}\int_0^C (1+t)^\beta dt d|\mu|(w,b)\\
&= (\beta+1)\frac{(1+C)^{\beta+1}-1}{\beta+1}\int_{\mathbb{S}^{d+1}} d|\mu|(w,b)\\
&= \Big((1+C)^{\beta+1}-1\Big)\int_{\mathbb{S}^{d+1}} d|\mu|(w,b)\\
&= \Big((1+C)^{\beta+1}-1\Big)\|\mu\|_{W_{\sigma_\beta},\Omega}.
\end{aligned}
$$

Taking the infimum over $\mu \in M^{\mathbb{S}^{d+1}}_{\sigma_{\beta+1},f}$ gives

$$\|f\|_{\mathcal{B}^\Omega_{\sigma_\beta}} \leqslant \Big((1+C)^{\beta+1}-1\Big)\|f\|_{\mathcal{B}^{\mathbb{S}^{d+1}}_{\sigma_\beta}}. \tag{202}$$

Equation (202) shows that eq. (194) holds. *Q.E.D.*

Theorem 3.1 can be seen as a stronger version of theorem 1 in [Li et al., 2020], and theorem 3.2 an extension thereof. The authors show that shallow neural networks with an activation function that satisfies

$$\int_{\mathbb{R}} \big|\partial^2\phi(x)\big|(1+|x|)dx < \infty \tag{203}$$

can be approximated well by a shallow neural network with ReLU as activation function. This requires that the second derivative exists, is bounded and decays fast enough. Theorem 3.1 and theorem 3.2 show that fast enough decay is not needed and in some cases boundedness isn't either. The benefit of their method is that they are able to compute the Rademacher complexity for the class with activation function ϕ very easily.

In conclusion, we see that the Barron spaces belonging to almost all activation functions ϕ in use, regardless of their parameter space Ω, embed in the Barron space with ReLU activation function on the unit ball.

3.5 Duality theorems

A (pre)dual of a vector space can help us find alternative formulations of problems, and may help us get a deeper understanding of the vector space itself. In this section we determine the dual and predual of V^Ω_ϕ. Both rely on the concept of the annihilator.

Definition 7 (Annihilator)**.** *The annihilator of a closed subset U of a Banach space Z with dual Z^* is given by*

$$U^\perp = \Big\{ z^* \in Z^* \;\Big|\; \forall z \in U : z^*(z) = 0 \Big\}. \tag{204}$$

3.5.1 Predual

Since V_ϕ^Ω is a quotient space of $rca(\Omega)$ we can determine the predual using the following lemma.

Lemma 3.2.1. *Let V be a Banach space and U some closed subset of V, then*

$$V^*/U^\perp \cong U^*. \tag{205}$$

Proof. Define

$$T : V^*/U^\perp \to U^*,\ [f] \mapsto \Big(v \mapsto T_{[f]}(v) = f(v)\Big). \tag{206}$$

T is linear by linearity of the functions in $V^*/U^\perp$. To show T is well defined, let $[f] = [f'] \in V^*/U^\perp$. Then $f - f' \in U^\perp$. Hence, for all $v \in U$

$$T_{[f]}(v) - T_{[f']}(v) = T_{[f-f']}(v) = (f - f')(v) = 0. \tag{207}$$

We prove that $\|T\| = 1$. For the smaller or equal, let $f \in U^*$. For every extension g of f to V^*, we have

$$\|[g]\|_{V^*/U^\perp} = \inf_{h\in U^\perp} \|g + h\|_{V^*} = \inf_{h\in U^\perp} \sup_{v\in V} |(g + h)(v)| \geqslant \inf_{h\in U^\perp} \sup_{v\in U} |(f + h)(v)| = \sup_{v\in U} |f(v)| = \|f\|_{U^*}. \tag{208}$$

However,

$$\big\|T_{[g]}\big\|_{U^*} = \big\|T_{[f]}\big\|_{U^*} = \|f\|_{U^*}. \tag{209}$$

Combined this gives

$$\big\|T_{[g]}\big\|_{U^*} \leqslant \|[g]\|_{V^*/U^\perp}. \tag{210}$$

For greater or equal, let $f \in U^*$. By Hahn Banach there exists a $g \in V^*$ such that for all $u \in U$

$$f(u) = g(u) \tag{211}$$

as well as

$$\|f\|_{U^*} = \|g\|_{V^*}. \tag{212}$$

Then

$$\begin{aligned} \|[g]\|_{V^*/U^\perp} &\leqslant \|g\|_{V^*} = \|f\|_{U^*} \\ \big\|T_{[f]}\big\|_{U^*} &= \|f\|_{U^*}. \end{aligned} \tag{213}$$

Combined this gives

$$\big\|T_{[g]}\big\|_{U^*} = \big\|T_{[f]}\big\|_{U^*} = \|f\|_{U^*} \geqslant \|[g]\|_{V^*/U^\perp}. \tag{214}$$

Q.E.D.

Lemma 3.2.1 shows us that we can identify the predual of V_ϕ with U if we can identify $M_{\phi,0}^\Omega$ with the annihilator of U.

Lemma 3.2.2. *If Ω is a compact set, then the annihilator of*

$$C^{\phi}_{rca(\Omega)}(\mathcal{X}) = \left\{ f \in C(\Omega) \;\middle|\; \exists \nu \in rca(\mathcal{X}) : f(w,b) = \int_{\mathcal{X}} \phi(\langle x|w\rangle + b) d\nu(x) \right\} \tag{215}$$

is $M^{\Omega}_{\phi,0}$ for $\phi \in C(\mathbb{R})$.

Proof. The annihilator of $C^{\phi}_{rca(\Omega)}(\mathcal{X})$ is given by

$$(C^{\phi}_{rca(\Omega)}(\mathcal{X}))^{\perp} = \left\{ \mu \in rca(\Omega) \;\middle|\; \forall f \in C^{\phi}_{rca(\Omega)}(\mathcal{X}) : \int_{\Omega} f(w,b) d\mu(w,b) = 0 \right\}. \tag{216}$$

We will show that $(C^{\phi}_{rca(\Omega)}(\mathcal{X}))^{\perp} = M^{\Omega}_{\phi,0}$ by showing inclusions. Observe that for $\mu \in rca(\Omega)$ and $\nu \in rca(\mathcal{X})$ by Fubini

$$\int_{\mathcal{X}} \int_{\Omega} \phi(\langle x|w\rangle + b) d\mu(w,b) d\nu(x) = \int_{\Omega} \int_{\mathcal{X}} \phi(\langle x|w\rangle + b) d\nu(x) d\mu(w,b), \tag{217}$$

which will be used in both directions of the proof.

First, we prove the right inclusion. Suppose $\mu \in (C^{\phi}_{rca(\Omega)}(\mathcal{X}))^{\perp}$, then for all $f \in C^{\phi}_{rca(\Omega)}(\mathcal{X})$ it holds that

$$\int_{\Omega} f(w,b) d\mu(w,b) = 0. \tag{218}$$

For each of these f there exists a $\nu \in rca(\mathcal{X})$ such that

$$f(w,b) = \int_{\mathcal{X}} \phi(\langle x|w\rangle + b) d\nu(x). \tag{219}$$

Hence,

$$0 = \int_{\Omega} \int_{\mathcal{X}} \phi(\langle x|w\rangle + b) d\nu(x) d\mu(w,b) = \int_{\mathcal{X}} \int_{\Omega} \phi(\langle x|w\rangle + b) d\mu(w,b) d\nu(x). \tag{220}$$

For this to hold for all ν, it must be that

$$\int_{\Omega} \phi(\langle x|w\rangle + b) d\mu(w,b) = 0. \tag{221}$$

This implies that $\mu \in M^{\Omega}_{\phi,0}$.

For the left inclusion, let $\mu \in M^{\Omega}_{\phi,0}$. This implies that

$$\int_{\Omega} \phi(\langle x|w\rangle + b) d\mu(w,b) = 0. \tag{222}$$

For each $f \in C^{\phi}_{rca(\Omega)}(\mathcal{X})$ there exists a $\nu \in rca(\mathcal{X})$ such that

$$f(w,b) = \int_{\mathcal{X}} \phi(\langle x|w\rangle + b) d\nu(x). \tag{223}$$

By Fubini

$$\int_\Omega f(w,b)d\mu(w,b) = \int_\Omega \int_\mathcal{X} \phi(\langle x|w\rangle + b)d\nu(x)d\mu(w,b) = \int_\mathcal{X} \int_\Omega \phi(\langle x|w\rangle + b)d\mu(w,b)d\nu(x) = 0. \quad (224)$$

This implies that $\mu \in (C^\phi_{rca(\Omega)}(\mathcal{X}))^\perp$. *Q.E.D.*

Theorem 3.3 (Bach predual)**.** *If Ω is a compact set, then a predual of V^Ω_ϕ is $C^\phi_{rca(\Omega)}(\mathcal{X})$.*

Proof. In proposition 3.3 it was proven that

$$V^\Omega_\phi \cong rca(\Omega)/M^\Omega_{\phi,0}. \quad (225)$$

The combination of lemma 3.2.2 and lemma 3.2.1 gives

$$rca(\Omega)/M^\Omega_{\phi,0} \cong (C^\phi_{rca(\Omega)}(\mathcal{X}))^*. \quad (226)$$

Combined this becomes

$$V^\Omega_\phi \cong (C^\phi_{rca(\Omega)}(\mathcal{X}))^*. \quad (227)$$

This implies that $C^\phi_{rca(\Omega)}(\mathcal{X})$ is a predual of V_ϕ. *Q.E.D.*

Consider a set of points $x_i \in X$. $C^\phi_{rca(\Omega)}(\mathcal{X})$ can be seen as all continuous functions created from these points. The Bach functions are dual functions of those. It is unclear what the implication of this is.

3.5.2 Dual

For the dual we use another lemma associated to quotient spaces.

Lemma 3.3.1. *Let V be a Banach space, V^* its dual and U some closed linear subspace of V, then*

$$(V/U)^* \cong U^\perp. \quad (228)$$

Proof. Let $f \in U^\perp$, and define

$$T : U^\perp \to (V/U)^*,\ f \mapsto \Big([v] \mapsto T_f([v]) = f(v)\Big) \quad (229)$$

for all $v \in Z$. Since $f \in V^*$, f is linear and so is T_f. To see that T_f is well defined, suppose that $[v] = [v']$. This means that $v - v' \in U$ and

$$0 = Tf([v - v']) = f(v) - f(v'). \quad (230)$$

Therefore, $f(v) = f(v')$.

Finally, we prove that $\|T\| = 1$. For the smaller or equal, let $[v] \in V/U$ such that $\|[v]\|_{V/U} \leqslant 1$. Fix an $\epsilon > 0$, $f \in U^\perp$, and recall that for all $u \in U$ we have that $f(u) = 0$. There exists a $y \in U$ such that $\|v - y\|_X \leqslant 1 + \epsilon$. Observe that

$$|T_f([x])| = |f(x)| = |f(x) - f(y)| = |f(x - y)| \leqslant \|f\|_{V*}\|x - y\|_V \leqslant \|f\|_{V*}(1 + \epsilon). \tag{231}$$

Hence,

$$\|T\| = \sup_{f \in U^\perp} \frac{\|T_f\|_{(V/U)*}}{\|f\|_{V*}} \leqslant 1 + \epsilon. \tag{232}$$

Since ϵ was arbitrary, $\|T\| \leqslant 1$.

For the larger or equal, let $[v] \in V/U$ such that $\|v\|_V \leqslant 1$. By Hahn-Banach there exists a function $f \in V^*$ such that $f(U) = \{0\}$, $f(v) = 1$ and $\|f\|_{X*} \leqslant 1$. Observe that

$$|T_f([v])| = |f(v)| = 1. \tag{233}$$

Hence, $\|T\| \geqslant 1$. *Q.E.D.*

Theorem 3.4 (Bach dual)**.** *Let Ω be compact, and let*

$$\begin{aligned} &T_1 : V_\phi^\Omega \to rca(\Omega)/M_{\phi,0}^\Omega,\ f \to M_f^\phi \\ &T_1^* : (rca(\Omega)/M_{\phi,0}^\Omega)^* \to (V_\phi^\Omega)^*,\ f^* \mapsto \Big(g \mapsto f^*(T_2 g)\Big) \\ &T_2 : (M_{\phi,0}^\Omega)^\perp \to (rca(\Omega)/M_{\phi,0}^\Omega)^*,\ f \mapsto \Big([v] \mapsto T_f([v]) = f(v)\Big). \end{aligned} \tag{234}$$

The dual of V_ϕ is isometrically isomorphic to $(M_{\phi,0}^\Omega)^\perp$ using the map

$$T_3 : (M_{\phi,0}^\Omega)^\perp \to (V_\phi^\Omega)^*,\ f \mapsto T_2(T_1^* f). \tag{235}$$

Proof. Since $rca(\Omega)/M_{\phi,0}^\Omega$ satisfies the requirements for lemma 3.3.1, it follows that T_2 is the isometric isomorphism between $\Big(rca(\Omega)/M_{\phi,0}^\Omega\Big)^*$ and $(M_{\phi,0}^\Omega)^\perp$. Hence, to show that T_3 is the isometric isomophism between $(V_\phi^\Omega)^*$ and $(M_{\phi,0}^\Omega)^\perp$, it is sufficient to show that T_1^* is an isometric isomophism between $(V_\phi^\Omega)^*$ and $(rca(\Omega)M_{\phi,0}^\Omega)^*$.

In proposition 3.3 it was established that T_1 is the isometric isomorphism between V_ϕ^Ω and $rca(\Omega)/M_{\phi,0}^\Omega$. T_1^* is the adjoint of T_1, and thus it is an isometric isomorphism between $(V_\phi^\Omega)^*$ and $\Big(rca(\Omega)/M_{\phi,0}^\Omega\Big)^*$ by the properties of adjoints [Rudin, 2006]. *Q.E.D.*

The annihilator of $M_{\phi,0}^\Omega$ is a subset of the dual of $rca(\Omega)$. This dual space, $rca(\Omega)^*$, is not a nice space. It is unclear whether $(M_{\phi,0}^\Omega)^\perp$ is a nice space, and what the implications are if it is.

4 Taylor and Relu

In this section we will show that the Barron spaces with (higher order) ReLU are strongly linked to Taylor expansions and in particular the integral remainder. We will show that this is true for certain functions from $\mathbb{R} \to \mathbb{R}$, but the strategy used does not generalise to functions from $\mathbb{R}^d \to \mathbb{R}$. However, applying the 1-dimensional Taylor on the exponential allows us to do an expansion for certain functions from $\mathbb{R}^d \to \mathbb{R}$ in the Fourier domain.

4.1 Single variable functions

In proposition 4.1 we have the conditions for the Taylor expansions of functions from $\mathbb{R} \to \mathbb{R}$ and the resulting shape of the expansion.

Proposition 4.1. *Let $k \in \mathbb{Z}_+$ and let $f : \mathbb{R} \to \mathbb{R}$ be a k times differentiable function on $a \in \mathbb{R}$ and $\partial^k f$ be absolutely continuous on the closed interval between a and x, then $f(x)$ can be represented equivalently as*

$$f(x) = \sum_{i=0}^{k} \frac{\partial^i f(a)}{i!}(x-a)^i + \int_a^x \frac{\partial^{(k+1)} f(t)}{k!}(x-t)^k dt. \tag{236}$$

Proof. First note that $\partial^{(m+1)} f$ exists for all $m \leqslant k$ as an $L^1([a,x])$ function, since $\partial^m f$ is absolutely continuous on the closed interval between a and x. The proof follows by induction.

Base case: Suppose $k = 0$. According to the fundamental theorem of Calculus

$$f(x) = f(a) + \int_a^x \partial^1 f(t) dt. \tag{237}$$

This shows that the base case satisfies the required equation.

Induction step: Assume $k \in \mathbb{Z}_+$ and $k > 0$ and that the statement is true for $k-1$. This means that

$$f(x) = \sum_{i=0}^{k-1} \frac{\partial^i f(a)}{i!}(x-a)^i + \int_a^x \frac{\partial^k f(t)}{(k-1)!}(x-t)^{k-1} dt \tag{238}$$

holds. Integration by parts on the integral gives

$$\int_a^x \frac{\partial^k f(t)}{(k-1)!}(x-t)^{k-1} dt = \frac{\partial^k f(a)}{k!}(x-a)^k + \int_a^x \frac{\partial^{k+1} f(t)}{k!}(x-t)^k dt. \tag{239}$$

Substitution of eq. (239) into eq. (238) gives

$$f(x) = \sum_{i=0}^{k} \frac{\partial^i f(a)}{i!}(x-a)^i + \int_a^x \frac{\partial^{(k+1)} f(t)}{k!}(x-t)^k dt. \tag{240}$$

This shows that the expansion holds for k as well, and thus for all $k \in \mathbb{Z}_+$. *Q.E.D.*

To link the Taylor expansion to ReLU we take a look at the integral form of the remainder. Using the identity

$$x^k = \sigma_k(x) + (-1)^k \sigma_k(-x) \quad \forall x \in \mathbb{R}^d \tag{241}$$

we can write the remainder as

$$\int_a^x \frac{\partial^{k+1} f(t)}{k!}(x-t)^k dt = \int_a^x \frac{\partial^{k+1} f(t)}{k!}\Big(\sigma_k(x-t) + (-1)^k \sigma_k(-x+t)\Big)dt. \tag{242}$$

Although this introduces the higher order ReLU, the limits of integration are dependent on x. For it to be written as a Barron function, we need the limits of integration to be independent of x. A slightly different approach gives us lemma 4.0.1. The idea is to construct a pair of higher order ReLU such that their integral over $[0, x]$ matches the original integral, and extend the integral to a fixed endpoint with the integrand zero on the extension.

Lemma 4.0.1. *Let $k \in \mathbb{Z}_+$, $z, c > 0$ such that $|z| \leqslant c$ and $f \in L^1([-c, c])$, then*

$$\int_0^z (z-u)^k f(u)du = \int_0^c \sigma_k(z-u)f(u) + (-1)^{k-1}\sigma_k(-z-u)f(-u)du. \tag{243}$$

Proof. Depending on the sign of z we can write the left hand side equivalently as

$$\int_0^z (z-u)^k f(u)du = \begin{cases} (-1)^{k-1}\int_0^c (-z-u)^k \mathbb{1}_{z<-u} f(-u)du & -c \leqslant z \leqslant 0 \\ \int_0^c (z-u)^k \mathbb{1}_{z>u} f(u)du & 0 \leqslant z \leqslant c \end{cases} \tag{244}$$

where the term $(-1)^{k-1}$ restores the sign for even k. Since both representations are zero in the domain of the other, we can add them to obtain

$$\int_0^z (z-u)^k f(u)du = \int_0^c (z-u)^k \mathbb{1}_{z>u} f(u) + (-1)^{k-1}(-z-u)^k \mathbb{1}_{z<-u} f(-u)du. \tag{245}$$

Note that

$$\begin{aligned} (z-u)^k \mathbb{1}_{z>u} &= \sigma_k(z-u), \\ (-z-u)^k \mathbb{1}_{z<-u} &= \sigma_k(-z-u). \end{aligned} \tag{246}$$

Substitution finishes the proof. *Q.E.D.*

The Taylor remainder can be written as a shallow neural network using lemma 4.0.1.

Theorem 4.1. *Let $k \in \mathbb{Z}_+$, $c \in \mathbb{R}_+$, $f : \mathbb{R} \to \mathbb{R}$ be a k times differentiable function on $[-c, c]$, and $\partial^k f$ be absolutely continuous on $[-c, c]$, then $f(x)$ can be represented equivalently as*

$$f(x) = \sum_{i=0}^{k} \frac{\partial^i f(a)}{i!}(x-a)^i + \frac{1}{k!}\int_0^c \sigma_k(x-t)\partial^{(k+1)} f(t) + (-1)^{k-1}\sigma_k(-x-t)\partial^{(k+1)} f(-t)dt \tag{247}$$

for all $x, a \in [-c, c]$.

Proof. According to proposition 4.1 we can write $f(x)$ as

$$f(x) = \sum_{i=0}^{k} \frac{\partial^i f(a)}{i!}(x-a)^i + \int_a^x \frac{\partial^{(k+1)} f(t)}{k!}(x-t)^k dt. \tag{248}$$

The integral can be written as

$$\int_a^x \frac{\partial^{(k+1)} f(t)}{k!}(x-t)^k dt = \frac{1}{k!}\int_0^c \sigma_k(x-t)\partial^{(k+1)} f(t) + (-1)^{k-1}\sigma_k(-x-t)\partial^{(k+1)} f(-t)dt \quad (249)$$

according to lemma 4.0.1. Substitution finishes the proof. *Q.E.D.*

Corollary 4.1.1. *Let $c \in \mathbb{R}_+$, $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function on $[-c, c]$, and $\partial^1 f$ be absolutely continuous on $[-c, c]$, then $f(x)$ can be represented equivalently as*

$$f(x) = f(a) + \frac{1}{2}\partial f(a)(x-a) + \frac{1}{2}\int_0^c \sigma(x-t)\partial^2 f(t) - \sigma(-x-t)\partial^2 f(-t)dt \quad (250)$$

for all $x, a \in [-c, c]$.

Corollary 4.1.1 shows that a shallow neural network with ReLU as activation function approximating a 1D function can be interpreted as a first order Taylor expansion with an approximation of the remainder.

4.2 Multivariate functions

When dealing with shallow neural networks though, we are interested in functions from $\mathbb{R}^d \to \mathbb{R}$ not just $\mathbb{R} \to \mathbb{R}$. In this section we will repeat the steps from the previous section for $d > 1$ to show that similar results do not hold for multivariate functions in the time domain: The Taylor expansion can be formulated for functions from $C^k(\mathbb{R}^d, \mathbb{R})$, but the remainder cannot be written as an infinitely wide neural network.

Proposition 4.2. *Let $k \in \mathbb{Z}_+$, $f : \mathbb{R}^d \to \mathbb{R}$ be a k times differentiable function on some ball B_a centered at $a \in \mathbb{R}^d$ and $\partial^{(k)} f$ be Lipschitz continuous on B_a, then $f(x)$ can be represented equivalently as*

$$f(x) = \sum_{|\alpha|=0}^{k} \frac{1}{\alpha!}\partial^\alpha f(a)(x-a)^\alpha + \sum_{|\alpha|=k+1} \frac{k+1}{\alpha!}(x-a)^\alpha \int_0^1 (1-t)^k \partial^\alpha f(a + t(x-a))dt, \quad (251)$$

where we have used the multi-index notation for α.

Proof. Consider $u(t) = a + t(x-a)$ and set $g(t) = f(u(t))$. $u(t)$ parametrizes the line between a and x. Observe that $g(0) = f(a)$ and $g(1) = f(x)$. Now we apply lemma 4.0.1 to g and get

$$f(x) = g(1) = \sum_{i=0}^{k} \frac{\partial^i g(0)}{i!} + \int_0^1 \frac{\partial^{(k+1)} g(t)}{k!}(1-t)^k dt. \quad (252)$$

The derivatives of g at t for $j \leqslant k+1$, which exist (as L^1 function for $k+1$) by virtue of $\partial^{(k)} f$ being Lipschitz continuous, are given by

$$\partial^j g(t) = \frac{\partial^j f(u(t))}{\partial t^j}$$

$$
\begin{aligned}
&= \frac{\partial^j f(a + t(x-a))}{\partial t^j} \\
&= \sum_{|\alpha|=j} \binom{j}{\alpha} \partial^\alpha f(a + t(x-a))(x-a)^\alpha.
\end{aligned}
$$

Since

$$
\frac{1}{j!}\binom{j}{\alpha} = \frac{1}{j!}\frac{j!}{\alpha!} = \frac{1}{\alpha!}, \tag{253}
$$

substitution of the derivatives of g into f finishes the proof. *Q.E.D.*

In the univariate case we used lemma 4.0.1 to rewrite the integral form of the remainder to the form of a shallow neural network. This heavily relied on the fact that the $\partial^k f(t)$ term was independent of x. In the multivariate case we have $\partial^\alpha f(a + t(x-a))$ instead. This is dependent on x, and thus lemma 4.0.1 cannot be used to rewrite the integral form of the remainder to the form of a shallow neural network. It is unclear whether there is another way to write the remainder as an infinitely wide neural network.

4.3 Fourier Expansion

In the time domain it is not clear how the remainder can be written as an infinitely wide neural network. However, it is possible to do so in the frequency domain. Recall that the inverse Fourier transform of $f : \mathbb{R}^d \to \mathbb{R}$ is given by

$$
f(x) = \int_{\mathbb{R}^d} e^{i\langle x|\xi\rangle} \hat{f}(\xi) d\xi \tag{254}
$$

when $\hat{f} \in L^1(\mathbb{R}^d)$ and $f \in L^1(\mathbb{R}^d)$, and that multiplying with $i\xi$ in the Fourier domain is equal to taking a derivative in the time domain. Proposition 4.1 gives the Taylor expansion for real valued functions. It is clear that this can be extended to complex valued functions, thus we can Taylor the exponential, even if the inputs are complex valued. This would suggest that, under smoothness assumptions of $f(x)$, the expression

$$
\int_{\mathbb{R}^d} \left(e^{i\langle x|\xi\rangle} - \sum_{k=0}^{s} \frac{i^k}{k!} \langle x|\xi\rangle^k \right) \hat{f}(\xi) d\xi \tag{255}
$$

exists, is finite, and can be approximated by integral(s) over (higher order) ReLUs. To make sure that all the Fourier terms are well-defined we require that $f \in \mathscr{F}_I^{s,1}$.

***Definition* 8.** *(Spectral Space) Let $s \in \mathbb{N}$. The spectral space $(\mathscr{F}_I^{s,1}, \|\cdot\|_{\mathscr{F}_I^{s,1}})$ is given*

$$
\begin{aligned}
\mathscr{F}_I^{s,1} &= \left\{ f : \mathbb{R}^d \to \mathbb{R} \;\middle|\; \|f\|_{\mathscr{F}_I^{s,1}} < \infty \right\} \\
\|f\|_{\mathscr{F}_I^{s,1}} &= \int_{\mathbb{R}^d} (1 + \|\xi\|_{\ell^1})^s \left| \hat{f}(\xi) \right| d\xi.
\end{aligned} \tag{256}
$$

Futhermore, set

$$
\|f\|_{\mathscr{F}^{s,1}} = \int_{\mathbb{R}^d} \|\xi\|_{\ell^1}^s \left| \hat{f}(\xi) \right| d\xi. \tag{257}
$$

To simplify the notation we mostly omit the ℓ^1 subscript from the norm.

Remark. *Since $\mathbb{R}^d$ is finite dimensional, it does not matter which $p \geqslant 1$ we use for the ℓ^p in the norm for which functions are in $\mathscr{F}_I^{s,1}$. Since we have several inner products between $x \in \mathcal{X}$ and $\xi \in \mathbb{R}^d$ in the integrals of this section, it is convenient to take $p = 1$ for the ℓ^p norm inside the spectral space definition.*

In section 6.2 we discuss what kind of functions are included in $\mathscr{F}_I^{s,1}$.

We start by taking the Taylor expansion of the exponential in the inverse Fourier transform of eq. (254).

Lemma 4.1.1. *Let $s \in \mathbb{N}$, then there exists an $R > 0$ such that*

$$e^{i\langle x|\xi\rangle} = \sum_{k=0}^{s} \frac{i^k}{k!} \langle x|\xi\rangle^k + \frac{i^{s+1}}{s!} \int_0^{\|\xi\|_1 R} \sigma_s(\langle x|\xi\rangle - u)e^{iu} + (-1)^{s-1}\sigma_s(-\langle x|\xi\rangle - u)e^{-iu}du \tag{258}$$

for every $(x, \xi) \in \mathcal{X} \times \mathbb{R}^d$.

Proof. Since $\mathcal{X}$ is bounded, there is a closed ball B of radius R such that $\mathcal{X} \subseteq B$. Using Hölder we have, for each $x \in \mathcal{X}$ and $\xi \in \mathbb{R}^d$, that

$$|\langle x|\xi\rangle| \leqslant \|x\|_{\ell^\infty} \|\xi\|_{\ell^1} \leqslant R\|\xi\|_{\ell^1}. \tag{259}$$

Hence, according to theorem 4.1 with $c = R\|\xi\|_1$ and $a = 0$, we have that

$$e^{iz} = \sum_{k=0}^{s} \frac{i^k}{k!} z^k + \frac{i^{s+1}}{s!} \int_0^{\|\xi\|_1 R} \sigma_s(\langle x|\xi\rangle - u)e^{iu} + (-1)^{s-1}\sigma_s(-\langle x|\xi\rangle - u)e^{-iu}du \tag{260}$$

for $z \in [-c, c]$. *Q.E.D.*

After taking the inverse Fourier transform of the polynomial part $\sum_{k=0}^{s} \frac{i^k}{k!} \langle x|\xi\rangle^k$, the polynomial part of the Taylor expansion in the time domain is retrieved.

Lemma 4.1.2. *Let $s \in \mathbb{N}$ and $f \in \mathscr{F}_I^{s,1}$ be sufficiently smooth, then*

$$\int_{\mathbb{R}^d} \sum_{k=0}^{s} \frac{i^k}{k!} \langle x|\xi\rangle^k e^{i\langle x|\xi\rangle} \hat{f}(\xi)d\xi = \sum_{|\beta|\leqslant s} \frac{1}{\beta!} \partial^\beta f(x) x^\beta \tag{261}$$

with β a multi-index.

Proof. The Fourier derivation identity in multi-index notation is given by

$$i^{|\beta|}\xi^\beta \hat{f}(\xi) = \widehat{\partial^\beta f}(\xi). \tag{262}$$

Furthermore, the k-th power of the inner product of x with y in multi-index notation is given by

$$\langle x|y\rangle^k = \left(\sum_{n=1}^{d} x_n y_n\right)^k = \sum_{|\alpha|=k} \binom{k}{\alpha} x^\alpha y^\alpha. \tag{263}$$

This means that for a fixed k

$$\frac{i^k}{k!}\langle x|\xi\rangle^k \hat{f}(\xi) = \frac{1}{k!}\langle x|i\xi\rangle^k \hat{f}(\xi) = \frac{1}{k!}\sum_{|\alpha|=k}\binom{k}{\alpha}x^\alpha (i\xi)^\alpha \hat{f}(\xi) = \sum_{|\alpha|=k}\frac{1}{\alpha!}\widehat{\partial^\alpha f}(\xi)x^\alpha. \tag{264}$$

Summing this over k from 0 to s and then taking the inverse Fourier transform gives the sought expression. *Q.E.D.*

The main theorem follows.

Theorem 4.2. *Fix $s \in \mathbb{N}$. If f is s times continuously differentiable on the ball $B \subseteq \mathcal{X}$ of radius R and $f \in \mathscr{F}_I^{s+1,1}$, then there exists a measure μ such that*

$$f(x) = \sum_{|\beta|\leqslant s}\frac{1}{\beta!}\partial^\beta f(0)x^\beta + \int_{\mathbb{S}^d \times [0,R]} \sigma_s(\langle x|w\rangle + b)d\mu(w,b) \tag{265}$$

for all $x \in B$.

Proof. f can be expressed using its inverse Fourier transform as

$$f(x) = \int_{\mathbb{R}^d} e^{i\langle x|\xi\rangle}\hat{f}(\xi)d\xi. \tag{266}$$

Using lemma 4.1.1 this becomes

$$f(x) = \int_{\mathbb{R}^d}\sum_{k=0}^{s}\frac{i^k}{k!}\langle x|\xi\rangle^k \hat{f}(\xi)d\xi + \frac{i^{s+1}}{s!}\int_{\mathbb{R}^d}\int_0^{\|\xi\|_1 R}\sigma_s(\langle x|\xi\rangle - u)e^{iu} + (-1)^{s-1}\sigma_s(-\langle x|\xi\rangle - u)e^{-iu}du\hat{f}(\xi)d\xi. \tag{267}$$

The first term on the right can be rewritten using lemma 4.1.2 such that

$$f(x) = \sum_{|\beta|\leqslant s}\frac{1}{\beta!}\partial^\beta f(0)x^\beta + \frac{i^{s+1}}{s!}\int_{\mathbb{R}^d}\int_0^{\|\xi\|_1 R}\sigma_s(\langle x|\xi\rangle - u)e^{iu} + (-1)^{s-1}\sigma_s(-\langle x|\xi\rangle - u)e^{-iu}du\hat{f}(\xi)d\xi. \tag{268}$$

Call the second term on the right $\tilde{f}$. To remove the dependence on ξ from the integral bounds of the inner integral of $\tilde{f}$, apply the change of coordinate

$$u = \|\xi\|t \qquad\qquad t \in [0,R].$$

As a result of this change the term corresponding to the bias depends on ξ.

$$\tilde{f}(x) = \frac{i^{s+1}}{s!}\int_{\mathbb{R}^d}\|\xi\|\int_0^R \sigma_s(\langle x|\xi\rangle - \|\xi\|t)e^{i\|\xi\|t} + (-1)^{s-1}\sigma_s(-\langle x|\xi\rangle - \|\xi\|t)e^{-i\|\xi\|t}dt\hat{f}(\xi)d\xi \tag{269}$$

Removing this using the homogeneity of σ_s results in

$$\tilde{f}(x) = \frac{i^{s+1}}{s!}\int_{\mathbb{R}^d}\|\xi\|^{s+1}\int_0^R \sigma_s(\left\langle x\middle|\frac{\xi}{\|\xi\|}\right\rangle - t)e^{i\|\xi\|t} + (-1)^{s-1}\sigma_s(-\left\langle x\middle|\frac{\xi}{\|\xi\|}\right\rangle - t)e^{-i\|\xi\|t}dt\hat{f}(\xi)d\xi. \tag{270}$$

Now the left hand side is real whilst the right hand side is complex. This means the imaginary part of the right hand side must be zero. To retrieve the format of the real part of the right hand side first rewrite $\hat{f}$ such that

$$\hat{f}(\xi)d\xi = e^{i\theta(\xi)}dF(\xi) \tag{271}$$

where F is a signed measure encoding the magnitude of $\hat{\xi}$, and $\theta(\xi)$ is the corresponding argument. After substitution observe that the real part is determined by the power of s.

$$\tilde{f}(x) = \frac{i^{s+1}}{s!}\int_{\mathbb{R}^d}\|\xi\|^{s+1}\int_0^R \sigma_s(\left\langle x\middle|\frac{\xi}{\|\xi\|}\right\rangle - t)e^{i(\|\xi\|t+\theta(\xi))} + (-1)^{s-1}\sigma_s(-\left\langle x\middle|\frac{\xi}{\|\xi\|}\right\rangle - t)e^{-i(\|\xi\|t-\theta(\xi))}dtdF(\xi) \tag{272}$$

In particular,

$$\tilde{f}(x) = \begin{cases} \int_{\mathbb{R}^d}\int_0^R (-1)^{\frac{s+1}{2}}\frac{\|\xi\|^{s+1}}{s!}\left(\sigma_s(\left\langle x\middle|\frac{\xi}{\|\xi\|}\right\rangle - t)\cos\left(\|\xi\|t+\theta(\xi)\right) + (-1)^{s-1}\sigma_s(-\left\langle x\middle|\frac{\xi}{\|\xi\|}\right\rangle - t)\cos\left(-\|\xi\|t+\theta(\xi)\right)\right)dtdF(\xi) & s \text{ odd}, \\ \int_{\mathbb{R}^d}\int_0^R (-1)^{\frac{s+2}{2}}\frac{\|\xi\|^{s+1}}{s!}\left(\sigma_s(\left\langle x\middle|\frac{\xi}{\|\xi\|}\right\rangle - t)\sin\left(\|\xi\|t+\theta(\xi)\right) + (-1)^{s-1}\sigma_s(-\left\langle x\middle|\frac{\xi}{\|\xi\|}\right\rangle - t)\sin\left(-\|\xi\|t+\theta(\xi)\right)\right)dtdF(\xi) & s \text{ even}. \end{cases}$$

When $\tilde{f}$ is split into two parts

$$\tilde{f}(x) = \begin{cases} \int_{\mathbb{R}^d}\int_0^R (-1)^{\frac{s+1}{2}}\frac{\|\xi\|^{s+1}}{s!}\sigma_s(\left\langle x\middle|\frac{\xi}{\|\xi\|}\right\rangle - t)\cos\left(\|\xi\|t+\theta(\xi)\right)dtdF(\xi) & \\ + \int_{\mathbb{R}^d}\int_0^R (-1)^{\frac{3s-1}{2}}\frac{\|\xi\|^{s+1}}{s!}\sigma_s(-\left\langle x\middle|\frac{\xi}{\|\xi\|}\right\rangle - t)\cos\left(-\|\xi\|t+\theta(\xi)\right)dtdF(\xi) & s \text{ odd} \\ \int_{\mathbb{R}^d}\int_0^R (-1)^{\frac{s+2}{2}}\frac{\|\xi\|^{s+1}}{s!}\sigma_s(\left\langle x\middle|\frac{\xi}{\|\xi\|}\right\rangle - t)\sin\left(\|\xi\|t+\theta(\xi)\right)dtdF(\xi) & \\ + \int_{\mathbb{R}^d}\int_0^R (-1)^{\frac{s+2}{2}}\frac{\|\xi\|^{s+1}}{s!}(-1)^{s-1}\sigma_s(-\left\langle x\middle|\frac{\xi}{\|\xi\|}\right\rangle - t)\sin\left(-\|\xi\|t+\theta(\xi)\right)dtdF(\xi) & s \text{ even} \end{cases}$$

it is clear that the measures defined by

$$d\mu_1(\xi,t) = \begin{cases} (-1)^{\frac{s+1}{2}}\frac{\|\xi\|^{s+1}}{s!}\cos\left(\|\xi\|t+\theta(\xi)\right)dtdF(\xi) & s \text{ odd} \\ (-1)^{\frac{s+2}{2}}\frac{\|\xi\|^{s+1}}{s!}\sin\left(\|\xi\|t+\theta(\xi)\right)dtdF(\xi) & s \text{ even} \end{cases}$$

$$d\mu_2(\xi,t) = \begin{cases} (-1)^{\frac{3s-1}{2}}\frac{\|\xi\|^{s+1}}{s!}\cos\left(-\|\xi\|t+\theta(\xi)\right)dtdF(\xi) & s \text{ odd} \\ (-1)^{\frac{3s}{2}}\frac{\|\xi\|^{s+1}}{s!}\sin\left(-\|\xi\|t+\theta(\xi)\right)dtdF(\xi) & s \text{ even} \end{cases}$$

allow for $\tilde{f}$ to be written as

$$\tilde{f}(x) = \int_{\mathbb{R}^d}\int_0^R \sigma_s(\left\langle x\middle|\frac{\xi}{\|\xi\|}\right\rangle - t)d\mu_1(\xi,t) + \int_{\mathbb{R}^d}\int_0^R \sigma_s(\left\langle x\middle|\frac{-\xi}{\|\xi\|}\right\rangle - t)d\mu_2(\xi,t). \tag{273}$$

By axial symmetry

$$\int_{\mathbb{R}^d}\int_0^R \sigma_s(\left\langle x\middle|\frac{-\xi}{\|\xi\|}\right\rangle - t)d\mu_2(\xi,t) = \int_{\mathbb{R}^d}\int_0^R \sigma_s(\left\langle x\middle|\frac{\xi}{\|\xi\|}\right\rangle - t)d\mu_2(-\xi,t), \tag{274}$$

and thus

$$\tilde{f}(x) = \int_{\mathbb{R}^d}\int_0^R \sigma_s(\left\langle x\middle|\frac{\xi}{\|\xi\|}\right\rangle - t)d\mu(w,b) \tag{275}$$

for

$$d\mu(\xi,t) = d\mu_1(\xi,t) + d\mu_2(-\xi,t). \tag{276}$$

Finally, let ν be the pushforward of μ along the map Θ given by

$$\Theta : \mathbb{R}^d \times [0, R] \to \mathbb{S}^d \times [0, R], \quad (\xi, t) \mapsto (\frac{\xi}{\|\xi\|}, t), \tag{277}$$

then

$$\tilde{f}(x) = \int_{\mathbb{S}^d} \int_0^R \sigma_s(\langle x|w\rangle + b) d\nu(w, b). \tag{278}$$

Substituting $\tilde{f}$ back into f gives the sought expression. *Q.E.D.*

The measure μ of theorem 4.2 satisfies

$$\|\mu\|_{rca} \leqslant \frac{2}{s!} \|f\|_{\mathscr{F}^{s+1,1}}, \tag{279}$$

and thus

$$\left\|\tilde{f}\right\|_{\mathcal{B}_{\sigma_s}^{S^d \times [0,R]}} \leqslant \int_{S^d \times [0,R]} (\|w\| + |b|)^s d\mu(w, b) = \int_{S^d \times [0,R]} (1 + b)^s d\mu(w, b) \leqslant 2 \frac{(1 + R)^s}{s!} \|f\|_{\mathscr{F}^{s+1,1}}. \tag{280}$$

This means that $\tilde{f}$, which represents the remainder of the Taylor series in the frequency domain, is a Barron function. Note that $\tilde{f}$ only approximates the remainder accurately on the specific ball of radius R. The upper bound for the Barron norm scales exponentially in s with base R. So there is a hefty pay off between the upper bound of the Barron norm and the radius of the ball on which the function is properly approximated.

On page 4 of [Klusowski and Barron, 2018] Barron and Klusowski discuss a similar result under stronger assumptions. They too show state $\tilde{f}$ can be written as a Barron function when $f \in \mathscr{F}^{s,1}$ and that $f - \tilde{f}$ is then a polynomial of order s. There are a few notable differences between what Barron and Klusowski have proven and what is proven here. Barron and Klusowski take $R_{\mathcal{X}} = 1$, formally prove their results for $s = 2$ and $s = 3$, and state that

$$\|\mu\|_{rca(\mathbb{S}^d \times [0,1])} \leqslant \|f\|_{\mathscr{F}^{s+1,1}} \tag{281}$$

for $s \in \mathbb{N}$. Since the norm they are interested in does not depend on $R_{\mathcal{X}}$, it was sufficient to consider the case $R_{\mathcal{X}} = 1$. The Barron norm does depend on the size of the weights and biases. Hence, the distinction between values of $R_{\mathcal{X}}$ becomes more important. Setting $R_{\mathcal{X}} = 1$ gives a nearly identical parameter space Ω, though. The most notable difference is the lack of the term $\frac{2}{s!}$ in the norm bound of μ. It is unclear where this distinction comes from.

A second paper discussing similar results is [Parhi and Nowak, 2021]. Parhi and Nowak state in [Parhi and Nowak, 2021,theorem 23] a similar result, derived using a very different method. Their method requires that f satisfies some growth condition instead of having finite $\mathscr{F}^{s,1}$ norm. It is unknown whether this growth is implied by the smoothness requirements or vice versa. Another difference is that their $f - \tilde{f}$ is a generic polynomial instead of the Taylor expansion of f near zero. Parhi and Nowak show that the shallow neural network with m parameters estimating $\tilde{f}$ can be interpreted as the m best hyperplanes in Fourier space to estimate $\tilde{f}$. Hence, the μ from theorem 4.2 can be interpreted as giving weights to hyperplanes in Fourier spaces, and f_m approximating the Taylor remainder is the best approximation using m Fourier hyperplanes.

4.4 Error bound and Approximation Theorem

In this section the bound for the remainder from theorem 4.2 will be derived. First, we provide a bound for the approximation error in the form of a direct approximation theorem, then we bound the Rademacher complexity, and finally we combine them to formulate a bound for the remainder.

Recall that up to this point we have used two formulations for the Bach and Barron spaces; the formulation with which we introduced them in which the functions take the form

$$f(x) = K_\phi^\Omega \mu(x) = \int_\Omega \phi(\langle x|w\rangle + b) d\mu(w,b), \tag{282}$$

and the tilde formulation in which the function take the form

$$f(x) = \tilde{K}_\phi^\Omega(a,\pi)(x) = \int_\Omega a(w,b)\phi(\langle x|w\rangle + b) d\pi(w,b). \tag{283}$$

To prove an error bound we use a third formulation. In this formulation the density a is added to π as a third parameter, i.e. we now consider function of the form

$$f(x) = \bar{K}_\phi^\Omega \bar{\pi}(x) := \int_{\mathbb{R}\times\Omega} \phi(\langle x|w\rangle + b) d\bar{\pi}(a,w,b). \tag{284}$$

The right hand side of eq. (284) is an expectation of $\phi(\langle x|w\rangle + b)$ with $(a,w,b) \sim \bar{\pi}$. A finite approximation of this would be

$$f_m(x) = \frac{1}{m}\sum_{i=1}^m a_i \phi(\langle x|w_i\rangle + b_i) \tag{285}$$

with $(a_i, w_i, b_i) \sim \bar{\pi}$. This is a shallow neural network with m neurons in the hidden layer, just with a factor $\frac{1}{m}$ in front. The prefactor does not limit the shallow neural networks we can consider, since for $m \in \mathbb{N}$ the prefactor $\frac{1}{m}$ can simply be combined with a_i into a new $\tilde{a}_i$. Approximating an expectation with an empirical average is a Monte Carlo process. These processes have the property that the error between the expectation and the empirical average scales with the inverse of the number of samples. Hence, we expect that the error between a Barron function and the best shallow neural network approximating it scales with $\frac{1}{m}$. Before we show that this is indeed the case, we will make it more rigorous that the formulations of eq. (283) and eq. (284) are equivalent for the cases we are considering.

***Definition* 9.** *Let ϕ be a fixed activation function, $\bar{W} : \mathbb{R}\times\Omega \to \mathbb{R}$ non-negative and set*

$$\bar{K}_\phi^\Omega : \mathbb{P}(\mathbb{R}\times\Omega) \to L^2(\mathcal{X},\rho),\ \pi \mapsto \left(x \mapsto \int_\Omega a\phi(\langle x|w\rangle + b) d\pi(a,w,b)\right)$$
$$\bar{M}_{\phi,f}^\Omega = \left\{\pi \in \mathbb{P}(\mathbb{R}\times\Omega) \;\middle|\; \forall x \in \mathcal{X}:\ f(x) = \bar{K}_\phi^\Omega \pi(x)\right\}.$$

The special case that $f(x) = 0$ is denoted $\bar{M}_{\phi,0}^\Omega$. The space $(\bar{\mathcal{N}}_{\phi,W}^\Omega, \|\cdot\|_{\bar{\mathcal{N}}_{\phi,W}^\Omega})$ given by

$$\bar{\mathcal{N}}_{\phi,W}^\Omega = \left\{f : \mathcal{X} \to \mathbb{R} \;\middle|\; \|f\|_{\bar{\mathcal{N}}_{\phi,W}^\Omega} < \infty\right\}$$

$$\|f\|_{\bar{\mathcal{N}}^{\Omega}_{\phi,W}} = \inf_{\pi \in \bar{M}^{\Omega}_{\phi,f}} \int_{\Omega} W(a,w,b) d\pi(w,b)$$

is called a bar infinitely wide neural network space. If $\bar{\mathcal{N}}^{\Omega}_{\phi,1}$ *is written, it is meant that* $W(a,w,b) = 1$ *for all* $(a,(w,b)) \in \mathbb{R} \times \Omega$.

Proposition 4.3. *Let* ϕ *be a fixed activation function and* $W : \mathbb{R} \times \Omega \to \mathbb{R}$ *a non-zero function which can be written as*

$$W(c,w,b) = |c| W(w,b) \tag{286}$$

for some non-negative weight function $W : \Omega \to \mathbb{R}$*, then* $\tilde{\mathcal{N}}^{\Omega}_{\phi,\tilde{W}} \simeq \bar{\mathcal{N}}^{\Omega}_{\phi,\tilde{W}}$.

Proof. Consider the two maps

$$T_{\mathbb{P}} : \mathbb{P}(\mathbb{R} \times \Omega) \to \Big((\Omega \to \mathbb{R}) \times \mathbb{P}(\Omega) \Big),\ \bar{\pi} \mapsto (\int_{\mathbb{R}} a d\bar{\pi}^{(w,b)}(a), \pi) \tag{287}$$

and

$$T^{\mathbb{P}} : \Big((\Omega \to \mathbb{R}) \times \mathbb{P}(\Omega) \Big) \to \mathbb{P}(\mathbb{R} \times \Omega),\ (a,\pi) \mapsto \delta_a \otimes \pi. \tag{288}$$

The first map, $T_{\mathbb{P}}$, splits the probability measure $\bar{\pi} \in \mathbb{P}(\mathbb{R} \times \Omega)$ into a measure $\pi \in \mathbb{P}(\Omega)$ and a integral which depends (w,b). The second map, $T_{\mathbb{P}}$, combines the density a and the probability measure π into a single product measure $\delta_a \otimes \pi$ by concentrating the density a using a Dirac measure.

If $(a,\pi) \in \tilde{M}^{\Omega}_{\phi,f}$ for $f \in \tilde{\mathcal{N}}^{\Omega}_{\phi,W}$, then

$$\bar{K}^{\Omega}_{\phi} T^{\mathbb{P}}(a,\pi) = \int_{\mathbb{R}\times\Omega} c\phi(\langle x|w\rangle + b) d(\delta_a \otimes \pi)(c,w,b) = \int_{\Omega} a(w,b)\phi(\langle x|w\rangle + b) d\pi(w,b) = f(x). \tag{289}$$

Hence, $T^{\mathbb{P}}(a,\pi) \in \bar{M}^{\Omega}_{\phi,f}$. It follows that

$$\|f\|_{\bar{\mathcal{N}}^{\Omega}_{\phi,\tilde{W}}} \leqslant \int_{\mathbb{R}\times\Omega} \tilde{W}(c,w,b) d(\delta_a \otimes \pi)(c,w,b) = \int_{\Omega} \tilde{W}(a(w,b),w,b) d\pi(w,b). \tag{290}$$

Taking the infimum over $(a,\pi) \in \tilde{M}^{\Omega}_{\phi,f}$ gives

$$\|f\|_{\bar{\mathcal{N}}^{\Omega}_{\phi,W}} \leqslant \|f\|_{\tilde{\mathcal{N}}^{\Omega}_{\phi,W}}. \tag{291}$$

On the other hand, if $\bar{\pi} \in \bar{M}^{\Omega}_{\phi,f}$ for $f \in \bar{\mathcal{N}}^{\Omega}_{\phi,\tilde{W}}$, then

$$\begin{aligned}
\tilde{K}^{\Omega}_{\phi} T_{\mathbb{P}} \bar{\pi} &= \int_{\Omega} \int_{\mathbb{R}} a d\bar{\pi}^{(w,b)}(a) \phi(\langle x|w\rangle + b) d\pi(w,b) \\
&= \int_{\Omega} \int_{\mathbb{R}} a\phi(\langle x|w\rangle + b) d\bar{\pi}^{(w,b)}(a) d\pi(w,b) \\
&= \int_{\mathbb{R}} \int_{\Omega} a\phi(\langle x|w\rangle + b) d\bar{\pi}^{(w,b)}(a) d\pi(w,b) \\
&= \int_{\mathbb{R}\times\Omega} a\phi(\langle x|w\rangle + b) d\bar{\pi}(a,w,b) \\
&= f(x)
\end{aligned} \tag{292}$$

by Fubini. Hence, $T_{\mathbb{P}}\bar{\pi} \in \tilde{M}^{\Omega}_{\phi,f}$. It follows that

$$\begin{aligned} \|f\|_{\tilde{\mathcal{N}}^{\Omega}_{\phi,W}} &\leqslant \int_{\Omega} \tilde{W}(\int_{\mathbb{R}} a d\bar{\pi}^{(w,b)}(a), w, b) d\pi(w,b) \\ &= \int_{\Omega} \left| \int_{\mathbb{R}} a d\bar{\pi}^{(w,b)}(a) \right| W(w,b) d\pi(w,b) \\ &\leqslant \int_{\Omega} \int_{\mathbb{R}} |a| W(w,b) d\bar{\pi}^{(w,b)}(a) d\pi(w,b) \\ &= \int_{\Omega} \tilde{W}(a,w,b) d\bar{\pi}(a,w,b). \end{aligned} \tag{293}$$

Taking the infimum over $\bar{\pi} \in \bar{M}^{\Omega}_{\phi,f}$ gives

$$\|f\|_{\tilde{\mathcal{N}}^{\Omega}_{\phi,W}} \leqslant \|f\|_{\bar{\mathcal{N}}^{\Omega}_{\phi,W}}. \tag{294}$$

Q.E.D.

Now that we have shown that the formulations of eq. (283) and eq. (284) are indeed equivalent, we can provide an approximation theorem for the Barron spaces with higher order ReLU as activation function. This proof follows a similar structure to that of theorem 5 in [E et al., 2021].

Theorem 4.3 (Direct Approximation Theorem)**.** *For every $f \in \mathcal{B}^{\Omega}_{\sigma_\alpha}$ and every $m \in \mathbb{N}$ there exists a shallow neural network*

$$f_m(x) = \frac{1}{m} \sum_{i=1}^{m} a_i \phi(\langle x | w_i \rangle + b_i) \tag{295}$$

such that

$$\|f - f_m\|^2_{L^2(X,\rho)} \leqslant 3 \max(1, R^{2\alpha}_{\mathcal{X}}) \frac{\|f\|^2_{\mathcal{B}}}{m} \tag{296}$$

as well as

$$\|f_m\|_{\mathcal{B}_\phi} \leqslant 2 ||f||_{\mathcal{B}}. \tag{297}$$

Proof. From proposition 3.2 it follows that it is sufficient to prove the statement for $\Omega = \mathbb{S}^{d+1}$. From proposition 3.5 and proposition 4.3 it follows that $f \in \mathcal{B}^{\Omega}_{\sigma_\alpha}$ if and only if $f \in \bar{\mathcal{N}}^{\Omega}_{\phi,\tilde{W}}$ with $\phi = \sigma_\alpha$ and

$$\tilde{W}(a,w,b) = |a| W_{\sigma_\alpha}(w,b). \tag{298}$$

Suppose $f \in \bar{\mathcal{N}}^{\Omega}_{\phi,\tilde{W}}$. There must be at least one $\pi \in \mathbb{P}(\mathbb{R} \times \Omega)$ such that

$$f(x) = \bar{K}^{\Omega}_{\phi} \bar{\pi}(x) = \int_{\mathbb{R} \times \Omega} \phi(\langle x | w \rangle + b) d\bar{\pi}(a,w,b). \tag{299}$$

Let

$$f_m(x) = \frac{1}{m} \sum_{i=1}^{m} a_i \phi(\langle x | w_i \rangle + b_i) \tag{300}$$

with $(a_i, w_i, b_i) \in \pi$. From the linearity of the expectation it follows that

$$\mathbb{E}_\pi[f_m(x)] = \frac{1}{m}\sum_{i=1}^m \mathbb{E}_\pi[a_1\phi(\langle x|w_1\rangle + b_1] = f(x). \tag{301}$$

Therefore, by the definition of the variance

$$\begin{aligned}
\mathbb{E}_\pi[(f_m(x - f(x)])^2] &= \mathbb{E}_\pi[(f_m(x - \mathbb{E}_\pi[f_m(x)])^2] \\
&= \operatorname*{Var}_\pi f_m \\
&= \operatorname*{Var}_\pi \frac{1}{m}\sum_{i=1}^m a_i\phi(\langle x|w_i\rangle + b_i) \\
&= \frac{1}{m^2}\sum_{i=1}^m \operatorname*{Var}_\pi a_i\phi(\langle x|w_i\rangle + b_i) \\
&= \frac{1}{m}\operatorname*{Var}_\pi a_1\phi(\langle x|w_1\rangle + b_1) \\
&= \frac{1}{m}\int_\Omega a^2\phi(\langle x|w\rangle + b)^2 d\pi(a,w,b) - \left(\int_\Omega a\phi(\langle x|w\rangle + b)d\pi(a,w,b)\right)^2 \\
&\leqslant \frac{1}{m}\int_\Omega a^2\phi(\langle x|w\rangle + b)^2 d\pi(a,w,b) \\
&\leqslant \frac{1}{m}\int_\Omega |a|^2(\|x\|\|w\| + |b|)^{2\alpha} d\pi(a,w,b) \\
&\leqslant \frac{1}{m}\max(1, \|x\|^{2\alpha})\int_\Omega |a|^2(\|w\| + |b|)^{2\alpha} d\pi(a,w,b) \\
&\leqslant \frac{1}{m}\max(1, \|x\|^{2\alpha})\int_\Omega |a|^2 d\pi(a,w,b).
\end{aligned}$$

Without the squares in the integral, the integral term would the bar version of the Barron semi-norm of π. From Cauchy Schwartz it follows that

$$\int_\Omega |a|^2 d\pi(a,w,b) \leqslant \left(\int_\Omega |a| d\pi(a,w,b)\right)^2. \tag{302}$$

Hence, after taking the infimum over $\pi \in \bar{M}^\Omega_{\phi,f}$ we get

$$\mathbb{E}_\pi[(f_m(x - f(x)])^2] \leqslant \max(1, \|x\|^{2\alpha})\frac{\|f\|^2_{\bar{\mathcal{N}}^\Omega_{\phi,\tilde{W}}}}{m}. \tag{303}$$

Using Fubini we get

$$\begin{aligned}
\mathbb{E}_\pi[\|f_m - f\|^2_{L^2(\mathcal{X},\rho)}] &= \mathbb{E}_{x\sim\rho}[\mathbb{E}_\pi[f_m(x) - f(x)]] \\
&\leqslant \mathbb{E}_{x\sim\rho}[\max(1, \|x\|^{2\alpha})\frac{\|f\|^2_{\bar{\mathcal{N}}^\Omega_{\phi,\tilde{W}}}}{m}] \\
&= \mathbb{E}_{x\sim\rho}[\max(1, \|x\|^{2\alpha})]\frac{\|f\|^2_{\bar{\mathcal{N}}^\Omega_{\phi,\tilde{W}}}}{m} \\
&\leqslant \max(1, R^{2\alpha}_{\mathcal{X}})\frac{\|f\|^2_{\bar{\mathcal{N}}^\Omega_{\phi,\tilde{W}}}}{m}.
\end{aligned} \tag{304}$$

At the same time we have, again by linearity of the expectation, that

$$\mathbb{E}_\pi[\|f_m\|_{\bar{\mathcal{N}}^\Omega_{\phi,\tilde{W}}}] = \|f\|_{\bar{\mathcal{N}}^\Omega_{\phi,\tilde{W}}}. \tag{305}$$

Both eq. (296) and eq. (297) are bounds of an expectation. We are interested in bounds for the terms inside the expectations. To show that there exists at least one f_m that achieves both bounds (up to a constant) simultaneously, we use Markov's inequality. Define two events

$$\mathcal{E}_1 = \left\{ \|f_m - f\|^2_{L^2(\mathcal{X},\rho)} < 3\max(1, R_{\mathcal{X}}^{2\alpha}) \frac{\|f\|^2_{\bar{\mathcal{N}}^\Omega_{\phi,\tilde{W}}}}{m} \right\} \tag{306}$$

and

$$\mathcal{E}_2 = \left\{ \|f_m\|_{\bar{\mathcal{N}}^\Omega_{\phi,\tilde{W}}} < 2\|f\|_{\bar{\mathcal{N}}^\Omega_{\phi,\tilde{W}}} \right\}. \tag{307}$$

By Markov's inequality

$$\begin{aligned} \mathbb{P}(\mathcal{E}_1) = 1 - \mathbb{P}(\neg\mathcal{E}_1) &\geqslant 1 - \frac{\mathbb{E}_\pi[\|f - f_m\|^2_{L^2(X,\rho)}]}{3\max\{1, R_{\mathcal{X}}^{2\alpha}\} \frac{\|f\|^2_{\bar{\mathcal{N}}^\Omega_{\phi,\tilde{W}}}}{m}} \geqslant \frac{2}{3} \\ \mathbb{P}(\mathcal{E}_2) = 1 - \mathbb{P}(\neg\mathcal{E}_2) &\geqslant 1 - \frac{\mathbb{E}_\pi[\|f_m\|_{\bar{\mathcal{N}}^\Omega_{\phi,\tilde{W}}}]}{2\|f\|_{\bar{\mathcal{N}}^\Omega_{\phi,\tilde{W}}}} \geqslant \frac{1}{2}. \end{aligned} \tag{308}$$

Therefore, the odds of these events happening simultaneously are

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) = \mathbb{P}(\mathcal{E}_1) + \mathbb{P}(\mathcal{E}_1) - 1 \geqslant \frac{2}{3} + \frac{1}{2} - 1 > 0. \tag{309}$$

Since this probability is strictly positive and $\mathcal{B}^\Omega_{\sigma_\alpha} \simeq \bar{\mathcal{N}}^\Omega_{\phi,\tilde{W}}$, there must be at least one f_m of the form of eq. (295) that satisfies both eq. (296) and eq. (297). *Q.E.D.*

Remark. *The proof of theorem 4.3 bounds* $\mathbb{E}_{x\sim\rho}[\max(1, \|x\|^{2\alpha})]$ *by* $\max(1, R_{\mathcal{X}}^{2\alpha})$. *Since* ρ *has finite second moment, this bound is not tight. However, in practice you are more likely to know* $R_{\mathcal{X}}$ *than the second moment of* ρ.

Theorem 4.3 provides a bound for the approximation error. Although we didn't need to go the extra mile to prove that f_m satisfies eq. (297) for a bound for this error, this bound will be useful later on when computing the bound for the estimation error. To bound the estimation error we need to compute a bound for the Rademacher complexity.

Proposition 4.4. *If*

$$\mathcal{B}_{\sigma_\alpha,Q,m} = \left\{ f_m \in \mathcal{B}^\Omega_{\sigma_\alpha} \;\middle|\; \|f_m\|_{\mathcal{B}m} := \frac{1}{m}\sum_{i=1}^m |a_i|(\|w_i\| + |b_i|)^\alpha \leqslant Q \right\}, \tag{310}$$

then

$$\operatorname{Rad} \mathcal{B}_{\sigma_\alpha,Q,m} \leqslant \alpha Q(1 + R_{\mathcal{X}}) \max\{1, R_{\mathcal{X}}^{\alpha-1}\} \sqrt{\frac{2\log(2d+2)}{n}}. \tag{311}$$

Proof. From proposition 3.2 it follows that it is sufficient to prove the statement for $\Omega = \mathbb{S}^{d+1}$.

By definition of the Rademacher complexity we get

$$\begin{aligned}
\operatorname{Rad} \mathcal{B}_{\sigma_\alpha,Q,m} &= \mathbb{E}_{S\sim\rho^n} \frac{1}{|S|} \mathbb{E}_\chi \Bigg[\sup_{\substack{f_m \in \mathcal{B}^\Omega_{\sigma_\alpha} \\ \|f_m\|_{\mathcal{B}m} \leqslant Q}} \sum_{k=1}^{|S|} \chi_k f_m(x_k) \Bigg] \\
&= \mathbb{E}_{S\sim\rho^n} \frac{1}{|S|} \mathbb{E}_\chi \Bigg[\sup_{\substack{f_m \in \mathcal{B}^\Omega_{\sigma_\alpha} \\ \|f_m\|_{\mathcal{B}m} \leqslant Q}} \sum_{k=1}^{|S|} \chi_k \sum_{i=1}^{m} a_i \sigma_\alpha(\langle x_k | w_i \rangle + b_i) \Bigg] \\
&\leqslant Q \mathbb{E}_{S\sim\rho^n} \frac{1}{|S|} \mathbb{E}_\chi \Bigg[\sup_{\|u\|+|v|\leqslant 1} \sum_{k=1}^{|S|} \chi_k \sigma_\alpha(\langle x_k | u \rangle + v) \Bigg].
\end{aligned}$$

On the interval $[-1, 1]$ we have that σ_α is α-Lipschitz. Since

$$\frac{|\langle x_k | u \rangle + v|}{\max\{1, R_{\mathcal{X}}\}} \leqslant 1,$$

we have

$$\begin{aligned}
&Q \mathbb{E}_{S\sim\rho^n} \frac{1}{|S|} \mathbb{E}_\chi \Bigg[\sup_{\|u\|+|v|\leqslant 1} \sum_{k=1}^{|S|} \chi_k \sigma_\alpha(\langle x_k | u \rangle + v) \Bigg] \\
&= Q \max\{1, R_{\mathcal{X}}^\alpha\} \mathbb{E}_{S\sim\rho^n} \frac{1}{|S|} \mathbb{E}_\chi \Bigg[\sup_{\|u\|+|v|\leqslant 1} \sum_{k=1}^{|S|} \chi_k \sigma_\alpha\Big(\frac{\langle x_k | u \rangle + v}{\max\{1, R_{\mathcal{X}}\}}\Big) \Bigg] \\
&\leqslant \alpha Q \max\{1, R_{\mathcal{X}}^\alpha\} \mathbb{E}_{S\sim\rho^n} \frac{1}{|S|} \mathbb{E}_\chi \Bigg[\sup_{\|u\|+|v|\leqslant 1} \sum_{k=1}^{|S|} \chi_k \frac{\langle x_k | u \rangle + v}{\max\{1, R_{\mathcal{X}}\}} \Bigg]
\end{aligned}$$

by lemma 26.9 of [Shalev-Shwartz and Ben-David, 2014] and homogeneity of σ_α. Lastly,

$$\begin{aligned}
&\alpha Q \max\{1, R_{\mathcal{X}}^\alpha\} \mathbb{E}_{S\sim\rho^n} \frac{1}{|S|} \mathbb{E}_\chi \Bigg[\sup_{\|u\|+|v|\leqslant 1} \sum_{k=1}^{|S|} \chi_k \frac{\langle x_k | u \rangle + v}{\max\{1, R_{\mathcal{X}}\}} \Bigg] \\
&= \alpha Q \max\{1, R_{\mathcal{X}}^{\alpha-1}\} \mathbb{E}_{S\sim\rho^n} \frac{1}{|S|} \mathbb{E}_\chi \Bigg[\sup_{\|u\|+|v|\leqslant 1} \sum_{k=1}^{|S|} \chi_k (\langle x_k | u \rangle + v) \Bigg] \\
&= \alpha Q \max\{1, R_{\mathcal{X}}^{\alpha-1}\} \mathbb{E}_{S\sim\rho^n} \frac{1}{|S|} \mathbb{E}_\chi \Bigg[\sup_{\|u\|+|v|\leqslant 1} \sum_{k=1}^{|S|} \chi_k (\Big\langle \begin{pmatrix} x_k \\ 1 \end{pmatrix} \Big| \begin{pmatrix} u \\ v \end{pmatrix} \Big\rangle) \Bigg] \\
&\leqslant \alpha Q (1 + R_{\mathcal{X}}) \max\{1, R_{\mathcal{X}}^{\alpha-1}\} \sqrt{\frac{2 \log(2(d+1))}{n}} \\
&= \alpha Q (1 + R_{\mathcal{X}}) \max\{1, R_{\mathcal{X}}^{\alpha-1}\} \sqrt{\frac{2 \log(2d+2)}{n}}
\end{aligned}$$

by lemma 26.11 of [Shalev-Shwartz and Ben-David, 2014]. *Q.E.D.*

Armed with a bound for the approximation error and a bound for the Rademacher complexity, we can compute an upper bound for the error between a Barron function and the best shallow neural network with m neurons optimised over a set of n samples approximating it.

Theorem 4.4. *For every $f \in \mathcal{B}^{\Omega}_{\sigma_\alpha}$ and every $m \in \mathbb{N}$ there exists a shallow neural network*

$$f_{m,S}(x) = \sum_{i=1}^{m} a_i \phi(\langle x|w_i\rangle + b_i) \tag{312}$$

that minimizes

$$\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - f_{m,S}(x_i))^2$$

such that for every $\delta > 0$ with probability at least $1-\delta$

$$\|f - f_{m,S}\|^2_{L^2(\mathcal{X},\rho)} \leqslant 3\max\{1, R_{\mathcal{X}}^{2\alpha-1}\}\|f\|^2_{\mathcal{B}_{\sigma_\alpha}}\left(\max\{1, R_{\mathcal{X}}\}\frac{1}{m} + 16\alpha(1+R_{\mathcal{X}})\sqrt{\frac{2log(2d+2)}{n}} + 6\max\{1, R_{\mathcal{X}}\}\sqrt{\frac{2\log(2/\delta)}{n}}\right)$$

over the sets of n data points x_i sampled from ρ.

Proof. From proposition 3.2 it follows that it is sufficient to prove the statement for $\Omega = \mathbb{S}^{d+1}$.

From proposition 2.1 and proposition 2.3 it follows that with probability at least $1-\delta$ over the sets of n data points from ρ the bound

$$\|f - f_{m,S}\|^2_{L^2(\mathcal{X},\rho)} \leqslant \|f - f_m\|^2_{L^2(\mathcal{X},\rho)} + 4L\operatorname{Rad}\mathcal{B}_{\sigma_\alpha} + 2M\sqrt{\frac{2\log(2/\delta)}{n}}, \tag{313}$$

holds, where f_m minimises $\|f - f_m\|^2_{L^2(\mathcal{X},\rho)}$, $f_{m,S}$ minimises $\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - f_{m,n}(x_i))^2$, L is the Lipschitz constant of ℓ and M is the biggest value for ℓ. From theorem 4.3 we know that there is an $\tilde{f}_m$ such that

$$\begin{aligned}\left\|f - \tilde{f}_m\right\|^2_{L^2(\mathcal{X},\rho)} &\leqslant 3\frac{\|f\|^2_{\mathcal{B}^{\Omega}_{\sigma_\alpha}}}{m}\\ \left\|\tilde{f}_m\right\|_{\mathcal{B}^{\Omega}_{\sigma_\alpha}} &\leqslant 2||f||_{\mathcal{B}^{\Omega}_{\sigma_\alpha}}.\end{aligned}$$

Hence, it is possible to restrict the space from $\mathcal{B}^{\Omega}_{\sigma_\alpha}$ to

$$\mathcal{B}_{\sigma_\alpha,Q,m} = \left\{f_m \in \mathcal{B}^{\Omega}_{\sigma_\alpha} \;\middle|\; \|f_m\|_{\mathcal{B}m} := \frac{1}{m}\sum_{i=1}^{m}|a_i|(\|w_i\| + |b_i|)^\alpha \leqslant Q\right\}, \tag{314}$$

with $Q = 2||f||_{\mathcal{B}^{\Omega}_{\sigma_\alpha}}$ without losing $\tilde{f}_m$. Recall that for $f \in \mathcal{B}^{\Omega}_{\sigma_\alpha}$

$$|f(x)| \leqslant \max\{1, R^\alpha\}||f||_{\mathcal{B}^{\Omega}_{\sigma_\alpha}}. \tag{315}$$

With this restriction we have

$$\|f - f_m\|^2_{L^2(\mathcal{X},\rho)} \leqslant \left\|f - \tilde{f}_m\right\|^2_{L^2(\mathcal{X},\rho)} \leqslant 3\frac{\|f\|^2_{\mathcal{B}}}{m}$$

$$M = \max_{\tilde{f}_m \in \mathcal{B}_{\sigma_\alpha,Q,m}} (f - \tilde{f}_m)^2 \leqslant \max\{1, R^{2\alpha}\}(Q^2 + ||f||^2_{\mathcal{B}^\Omega_{\sigma_\alpha}} + 2Q||f||_{\mathcal{B}^\Omega_{\sigma_\alpha}}) = 9\max\{1, R^{2\alpha}\}||f||^2_{\mathcal{B}^\Omega_{\sigma_\alpha}}$$

$$\operatorname{Rad}\mathcal{B}_{\sigma_\alpha,Q,m} \leqslant 2\alpha||f||_{\mathcal{B}^\Omega_{\sigma_\alpha}}(1 + R_{\mathcal{X}})\max\{1, R^{\alpha-1}_{\mathcal{X}}\}\sqrt{\frac{2\log 2(d+1)}{n}}$$

and for $\tilde{f}_m, \tilde{g}_m \in \mathcal{B}_{\sigma_\alpha,Q,m}(\Omega)$ that

$$\begin{aligned}
\Big|(\tilde{f}_m(x) - f(x))^2 - (\tilde{g}_m(x) - f(x))^2\Big| &= \Big|(\tilde{f}_m(x) - f(x)) + (\tilde{g}_m(x) - f(x))\Big|\Big|(\tilde{f}_m(x) - f(x)) - (\tilde{g}_m(x) - f(x))\Big| \\
&= \Big|\tilde{f}_m(x) + \tilde{g}_m(x) - 2f(x)\Big|\Big|\tilde{f}_m(x) - \tilde{g}_m(x)\Big| \\
&\leqslant 2\max\{1, R^\alpha\}(Q + ||f||_{\mathcal{B}^\Omega_{\sigma_\alpha}})\Big|\tilde{f}_m(x) - \tilde{g}_m(x)\Big| \\
&\leqslant 6\max\{1, R^\alpha\}||f||_{\mathcal{B}^\Omega_{\sigma_\alpha}}\Big|\tilde{f}_m(x) - \tilde{g}_m(x)\Big| \\
&= L\Big|\tilde{f}_m(x) - \tilde{g}_m(x)\Big|,
\end{aligned}$$

where $L = 6\max\{1, R^\alpha\}||f||_{\mathcal{B}^\Omega_{\sigma_\alpha}}$. Thus, with probability at least $1 - \delta$ over the sets of n data points from $\mathcal{X}$ sampled according to ρ

$$\begin{aligned}
\|f - f_{m,n}\|^2_{L^2(\mathcal{X},\rho)} \leqslant\ & 3\max\{1, R^{2\alpha}\}\frac{\|f\|^2_{\mathcal{B}^\Omega_{\sigma_\alpha}}}{m} + 48\alpha||f||^2_{\mathcal{B}^\Omega_{\sigma_\alpha}}(1 + R_{\mathcal{X}})\max\{1, R^{2\alpha-1}_{\mathcal{X}}\}\sqrt{\frac{2\log 2(d+1)}{n}} \\
&+ 18\max\{1, R^{2\alpha}\}||f||^2_{\mathcal{B}^\Omega_{\sigma_\alpha}}\sqrt{\frac{2\log(2/\delta)}{n}}.
\end{aligned}$$

Factoring out common terms finishes the proof. *Q.E.D.*

Corollary 4.4.1. *Suppose $\tilde{f}_s$ is the order s remainder of f from theorem 4.2, then for every $m \in \mathbb{N}$ there exists a shallow neural network*

$$f_{m,S}(x) = \sum_{i=1}^m a_i\phi(\langle x|w_i\rangle + b_i) \tag{316}$$

that minimises

$$\frac{1}{n}\sum_{i=1}^n(\tilde{f}_s(x_i) - f_{m,S}(x_i))^2 \tag{317}$$

such that for every $\delta > 0$ with probability at least $1 - \delta$

$$\begin{aligned}
\|f - f_{m,S}\|^2_{L^2(\mathcal{X},\rho)} \leqslant\ & 12\max\{1, R^{2s-1}_{\mathcal{X}}\}\frac{(1 + R_{\mathcal{X}})^{2s}}{(s!)^2}\|f\|^2_{\mathscr{F}^{s+1,1}} \\
&\left(\max\{1, R_{\mathcal{X}}\}\frac{1}{m} + 16s(1 + R_{\mathcal{X}})\sqrt{\frac{2log(2d+2)}{n}} + 6\max\{1, R_{\mathcal{X}}\}\sqrt{\frac{2\log(2/\delta)}{n}}\right)
\end{aligned} \tag{318}$$

over the sets of n data points x_i sampled from ρ.

In both theorem 4.4 and corollary 4.4.1 we see that we have bounds of the form

$$\|f - f_{m,S}\|^2_{L^2(\mathcal{X},\rho)} \leqslant \mathcal{O}(\frac{1}{m} + \sqrt{\frac{1}{n}}). \tag{319}$$

This bound has just one term explicitly depending on d, and both $\frac{1}{m}$ and $\sqrt{\frac{1}{n}}$ are independent of d. This suggest that the Barron spaces with higher order ReLU do not suffer from the curse of dimensionality. However, just like before we also have that there is a hefty pay off between the upper bound and the radius of the ball on which we try to estimate the Barron function.

5 Numerics

We started this work with a discussion of various errors. Later, in section 4.3 we showed that the remainder of a Taylor expansion can be written using a Barron function, and provided an error bound. In this section we compare this bound to what can be realised in practice. The function whose remainder we are going to estimate is the function

$$f_g : \mathbb{R} \to \mathbb{R},\ x \mapsto e^{-\frac{1}{2}x^2}. \tag{320}$$

This function is (up to a constant) the one dimensional Gaussian with mean 0 and standard deviation 1. This makes its Fourier transform easy to compute, and in section 5.1 we show that $f_g \in \mathscr{F}_I^{s,1}$. It is also a smooth function, which means that it satisfies the requirements of theorem 4.2. In section 5.1 we also compute the upper bound analytically. In section 5.2 we describe the experiment based on this upper bound. Afterwards, the results of the experiment are presented and discussed in section 5.3.

The code for the experiment can be found on github/TJHeeringa/thesis.

5.1 Error bounds

It is clear that f_g is smooth and decays to zero at infinity. This means that $f_g \in C_0^\infty(\mathbb{R}^d)$. It is not immediately clear that $f_g \in \mathscr{F}_I^{s,1}$. To show this we need the Gamma function Γ given by

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt \tag{321}$$

for all $z \in \mathbb{C}$ with positive real part. This Γ function satisfies

$$\Gamma(n+1) = n! \tag{322}$$

for all $n \in \mathbb{N}$.

Proposition 5.1. *The Fourier transform of f_g is given by*

$$\hat{f}_g(\xi) = \sqrt{2\pi}e^{-\frac{1}{2}\xi^2},$$

and its $\mathscr{F}^{s,1}$ norm as function of s is

$$\|f_g\|_{\mathscr{F}^{s,1}} = 2^{\frac{s+1}{2}}\sqrt{2\pi}\Gamma(\frac{s+1}{2}).$$

Proof. The Fourier transform of f_g is given by

$$\hat{f}_g(\xi) = \int_{\mathbb{R}} f(x) e^{i\xi x} d\xi = \int_{\mathbb{R}} e^{-\frac{1}{2}x^2} e^{-i\xi x} d\xi.$$

Taking the derivative with respect to ξ gives

$$\partial_\xi \hat{f}_g(\xi) = -i \int_{\mathbb{R}} x e^{-\frac{1}{2}x^2} e^{-i\xi x} d\xi.$$

Using integration by parts results in

$$\partial_\xi \hat{f}_g(\xi) = \xi \int_{\mathbb{R}} e^{-\frac{1}{2}x^2} e^{-i\xi x} d\xi = \xi \hat{f}_g(\xi)$$

The general solution to this ODE is

$$\hat{f}_g(\xi) = c e^{-\frac{1}{2}\xi^2}.$$

for some $c \in \mathbb{R}$. Since

$$\hat{f}_g(0) = \int_{\mathbb{R}} e^{-\frac{1}{2}x^2} d\xi = \sqrt{2\pi},$$

it must be that $c = \sqrt{2\pi}$, and thus

$$\hat{f}_g(\xi) = \sqrt{2\pi} e^{-\frac{1}{2}\xi^2}.$$

Observe that $\hat{f}_g$ is a again a scaled Gaussian with zero mean. Since $\left|\hat{f}_g(\xi)\right| |\xi|^s$ is an even function of ξ and the interval of integration is symmetric about zero, this means

$$\|f_g\|_{\mathscr{F}^{s,1}} = \int_{\mathbb{R}} \left|\hat{f}_g(\xi)\right| |\xi|^s d\xi = 2 \int_0^\infty \hat{f}_g(\xi) \xi^s d\xi = 2\sqrt{2\pi} \int_0^\infty e^{-\frac{1}{2}\xi^2} \xi^s d\xi.$$

Let $\zeta = \frac{1}{2}\xi^2$, then

$$\|f_g\|_{\mathscr{F}^{s,1}} = 2^{\frac{s+1}{2}} \sqrt{2\pi} \int_0^\infty e^{-\zeta} \zeta^{\frac{s-1}{2}} d\zeta.$$

Recall that

$$\Gamma(k+1) = \int_0^\infty y^k e^{-y} dy,$$

which leads to

$$\|f_g\|_{\mathscr{F}^{s,1}} = 2^{\frac{s+1}{2}} \sqrt{2\pi} \Gamma(\frac{s+1}{2}).$$

Q.E.D.

Since f_g is smooth, we can Taylor it.

Proposition 5.2. *The Taylor remainder of order s for f_g is given by*

$$\tilde{f}_s = e^{-\frac{1}{2}x^2} - \sum_{k=0}^{\lfloor \frac{s}{2} \rfloor} \frac{(-1)^k x^{2k}}{2^k k!}. \tag{323}$$

Proof. Recall that

$$e^y = \sum_{k=0}^{\infty} \frac{y^k}{k!}$$

for all $y \in \mathbb{R}$. Hence,

$$f_g(x) = e^{-\frac{1}{2}x^2} = \sum_{k=0}^{\infty} \frac{(-\frac{1}{2}x^2)^k}{k!} = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{2^k k!}.$$

This can be written as a Taylor series up to order s by splitting on the x^s term.

$$f_g(x) = \sum_{k=0}^{\lfloor \frac{s}{2} \rfloor} \frac{(-1)^k x^{2k}}{2^k k!} + \underbrace{\sum_{k=\lceil \frac{s+1}{2} \rceil}^{\infty} \frac{(-1)^k x^{2k}}{2^k k!}}_{\text{remainder}}.$$

Hence, the remainder of f_g to order s, which we denote $\tilde{f}_s$, is given by

$$\tilde{f}_s = e^{-\frac{1}{2}x^2} - \sum_{k=0}^{\lfloor \frac{s}{2} \rfloor} \frac{(-1)^k x^{2k}}{2^k k!}.$$

Q.E.D.

Since $f_g \in C_0^\infty(\mathbb{R})$ and, according to proposition 5.1, $f_g \in \mathscr{F}^{s,1}$ for all $s \in \mathbb{N}$, it follows that f_g satisfies the conditions for theorem 4.2 for all s and all closed balls of finite radius. The associated upper bound of corollary 4.4.1 consists of three terms. These are given by

$$\begin{aligned}
\mathcal{E}_m(s;R) &:= 12 \max\{1, R_{\mathcal{X}}^{2s}\} \frac{(1+R)^{2s}}{(s!)^2} \|f\|^2_{\mathscr{F}^{s+1,1}} \\
\mathcal{E}_{d,n}(s;R) &:= 192 s \max\{1, R^{2s-1}\} \frac{(1+R_{\mathcal{X}})^{2s+1}}{(s!)^2} \sqrt{2log(2d+2))} \|f\|^2_{\mathscr{F}^{s+1,1}} \\
\mathcal{E}_{\delta,n}(s;R) &:= 72 \max\{1, R^{2s}\} \frac{(1+R_{\mathcal{X}})^{2s}}{(s!)^2} \sqrt{2\log(2/\delta)} \|f\|^2_{\mathscr{F}^{s+1,1}}
\end{aligned}$$

such that

$$\left\| \tilde{f}_s - f_{m,S} \right\|^2_{L^2(\mathcal{X},\rho)} \leqslant \mathcal{E}_m(s;R) \frac{1}{m} + \mathcal{E}_{d,n}(s;R) \frac{1}{\sqrt{n}} + \mathcal{E}_{\delta,n}(s;R) \frac{1}{\sqrt{n}}, \tag{324}$$

where we recall that $\tilde{f}_s$ is the remainder of the Taylor expansion of order s, $f_{m,S}$ is the best shallow neural network with m neurons in the hidden layer for the n samples in S, d is the dimension, s is the order of the Taylor expansion, R is the radius of the closed ball, and $1-\delta$ is the probability that

the bound is satisfied. To investigate how these bounds behave for increases s, we compute them for several radii. Suppose $\delta = 0.01$, then for $R = 1$ and $R = 3$ these terms become

$$\begin{aligned}
\mathcal{E}_m(s;3) &= 12\pi \frac{2^{5s+2}3^{2s}}{(s!)^2}\left(\Gamma(\frac{s+2}{2})\right)^2 & \mathcal{E}_m(s;1) &= 12\pi \frac{2^{3s+2}}{(s!)^2}\left(\Gamma(\frac{s+2}{2})\right)^2 \\
\mathcal{E}_{d,n}(s;3) &= 192\pi s \frac{2^{5s+4}3^{2s-1}}{(s!)^2}\sqrt{2log(4)}\left(\Gamma(\frac{s+2}{2})\right)^2 & \mathcal{E}_{d,n}(s;1) &= 192\pi s \frac{2^{3s+3}}{(s!)^2}\sqrt{2log(4)}\left(\Gamma(\frac{s+2}{2})\right)^2 \\
\mathcal{E}_{\delta,n}(s;3) &= 72\pi \frac{2^{5s+2}3^{2s}}{(s!)^2}\sqrt{2\log(200)}\left(\Gamma(\frac{s+2}{2})\right)^2 & \mathcal{E}_{\delta,n}(s;1) &= 72\pi \frac{2^{3s+5}}{(s!)^2}\sqrt{2\log(200)}\left(\Gamma(\frac{s+2}{2})\right)^2,
\end{aligned} \tag{325}$$

where we have that proposition 5.1 tells us that

$$\|f_g\|^2_{\mathscr{F}^{s,1}} = 2^{s+2}\pi\left(\Gamma(\frac{s+1}{2})\right)^2.$$

These terms have been plotted against s in fig. 2. At first the terms exponential in s increase faster, causing the error terms to increase. Later, the factorial takes over, causing the error terms to decrease. For higher R it takes the factorial longer to take over. The point where the factorial is sufficiently large to push the error below the error at $s = 1$, is the first s for which the solid line in below the dashed line. If these terms represented the true error instead of an upper bound thereof, then taking a higher order ReLU with an s between $s = 1$ and the crossover point would give worse results than just using a higher order ReLU with $s = 1$. When $R = 1$, this would be for values of $s \in \{2, \ldots, 12\}$.

5.2 Methodology

The bound of eq. (324) is a worst case bound. It is therefore unlikely that we achieve the bound, and we will likely stay well below it. This means that an experiment to see if we quantitatively match the behaviour does not make much sense. However, eq. (324) is an affine function of $\frac{1}{m}$ when n is fixed, and it is an affine function of $\frac{1}{\sqrt{n}}$ when m is fixed. This is qualitative behaviour we can test. We will now describe a numerical experiment to do so.

5.2.0.1 Data set

Theorem 4.2 says that we there is a Barron function describing the remainder of f_g that is valid up to a closed ball of radius R. f_g satisfies $f_g(0) = 1$ and $f_g(3) \approx 10^{-4}$. This means f_g is close to zero outside the interval $[-3, 3]$. At the same time fig. 1 suggests that we need a very high s to get an upper bound below that of $s = 1$ when $R = 3$. On the other hand, we need only $s = 13$ when $R = 1$. Hence, we will execute our experiment twice; once for $R = 1$ and once for $R = 3$. For each experiment we will take a closed ball B_R of radius R around the origin as the domain on which we approximate (the remainder of) f_g. We will sample $x \in B_R$ uniformly.

Since f_g is smooth and $f_g \in \mathscr{F}_I^{s,1}$ for all $s \in \mathbb{N}$, we could estimate the remainder for any $s \in \mathbb{N}$. We till take $s = 10$. This is the number of colours supported by default by Matplotlib. It is possible to

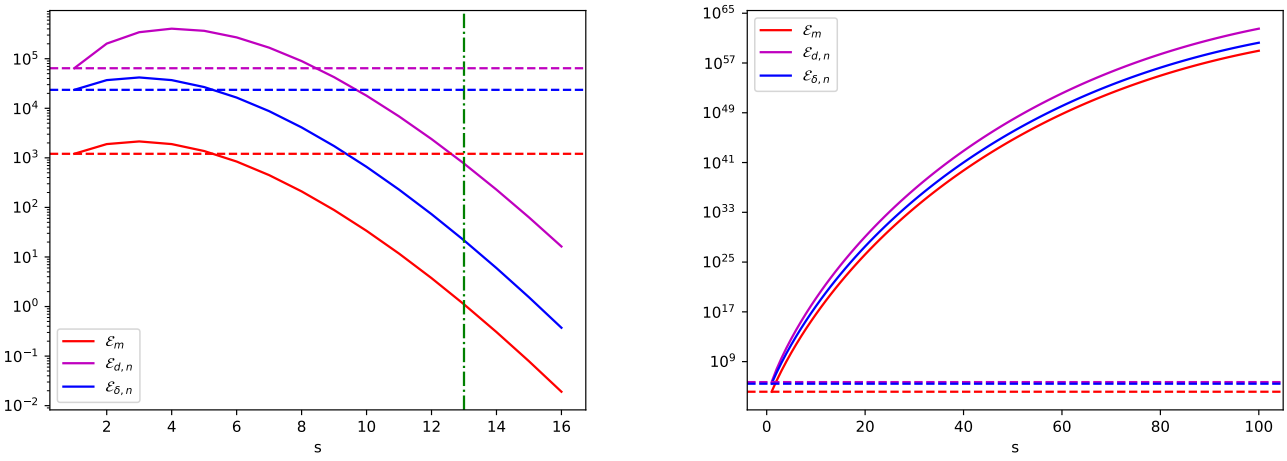


Figure 2: The solid lines are the terms of eq. (325) plotted against s. On the left the three terms corresponding to $R = 1$ have been plotted, whereas on the the right the three terms corresponding to $R = 3$. The dashed lines are placed at the values of the bounds for $s = 1$. When the solid line goes under the dashed line of the same colour, s is high enough to decrease the corresponding bound belong its size for $s = 1$. On the left this happens at $s = 13$, the location of the green dashdot line. On the right this does not happen within the plot.

increase the number of colours by defining a custom colour palette. However, plotting more than 10 lines on a plot, makes the plots hard to read. At the same time, $s = 10$ should be large enough to uncover the pattern.

Neural networks have the universal approximation property, so we can also directly estimate the function instead of estimating the remainder. Note that for these we have no explicit bound for the higher order ReLU.

With this we construct the data sets as $(x, \tilde{f}_s)$ tuples and (x, f_g) tuples for the remainder and full approximation experiments.

5.2.0.2 Training strategy

The data sets consist of labelled tuples, so we can train using supervised learning. We will use a shallow neural network with m neurons in the hidden layer for estimation. The order s of the shallow neural network will be the same as the order of the Taylor expansion for the $(x, \tilde{f}_s)$ data set. We will vary the order from $s = 1$ to $s = 10$ for the set (x, f_g). Since the available computation power is limited, we will vary m with n fixed at $n = 500$ from $m = 100$ to $m = 1000$ and vary n with m fixed at $m = 500$ from $n = 100$ to $n = 1000$. We will train the neural networks for 500 epochs with the Adam optimiser and batch size 32.

	overparametrized	underparametrized
Epochs	500	500
Batch size	32	32
Optimizer	Adam	Adam

Table 1: Parameters used in the numerical tests of this section.

5.2.0.3 Computing the bound

After training we will have 200 neural networks for each R. Each of these neural networks is a function $f_{m,S} : \mathbb{R} \to \mathbb{R}$. Since we know the function f_g and the remainder $\tilde{f}_s$ on the entire domain, we can compute both

$$\|f_g - f_{m,S}\|^2_{L^2(B_R,\rho)} \tag{326}$$

and

$$\left\|\tilde{f}_s - f_{m,S}\right\|^2_{L^2(B_R,\rho)} \tag{327}$$

using numerical integration. Note that this numerical integration introduces a small numerical error. This numerical error should be small enough to not distort the results. We will use the integration method 'scipy.quad'. This method gives back the value of the integral and a bound for the made numerical error, which we can use to verify whether the errors are indeed small. We run 'scipy.quad' with the option 'epsabs=1e-6' except for when we estimate eq. (326) on $R = 1$. This lowers the absolute error tolerance from the default of 1.98e-8 to 1e-6. This increases the speed but lowers the accuracy of 'scipy.quad'. Based on an initial test with Riemann sums, lowering the accuracy by setting the option 'epsabs=1e-6' does not impact the results of most cases. The exception to this is when we estimate eq. (327) on $R = 1$. Hence, in that case we still use the default settings.

5.3 Results

We will now discuss the results of the experiments.

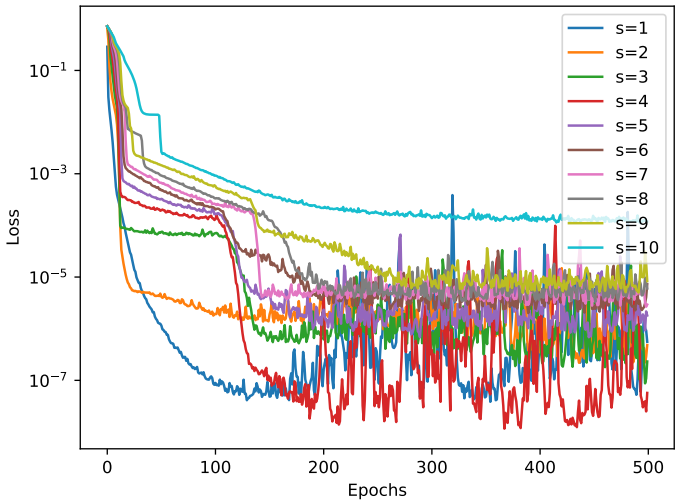
First, we take a look at the losses during training when trying to approximate f_g. The losses are plotted in fig. 3. In (a) and (b) we see the losses for the underparametrized case. We see in both that the loss goes down relatively smoothly and then turns noisy. This is seen more often in training using stochastic methods, and is due to the order that the data arrives in. The general pattern is that higher order ReLU with lower s achieve a lower loss than the those with higher s. In (a) this pattern does not hold for the lower s; there the order from bottom to top in later epochs is $s = 4, s = 1, s = 3, s = 2$. When we plot several of the final networks from the underparametrized case with radius $R = 3$ in fig. 4, we see that the networks with s larger struggle more with fitting the edges. A possible explanation is that a small change in a parameter leads to way bigger changes due to the high powers involved for higher s. This means that the tolerances on good parameters are lower, and suitable parameters may not be found at all. When we go to the overparametrized cases in (c) and (d), we see far less noisy behaviour at the end. This is expected. Since we are in the overparametrized regime, we should be able to find the parameters that drive the error to zero. This is best exemplified by $s = 1$ in (c). It does not have to go directly to zero, because the training can

get into local optima. This is best exemplified by $s = 10$, which jumps from local optimum to local optimum but continues trending to zero. In (d) we still expect the loss to go to zero, but except for $s = 1$ most seem to have gotten stuck in a local minimum.

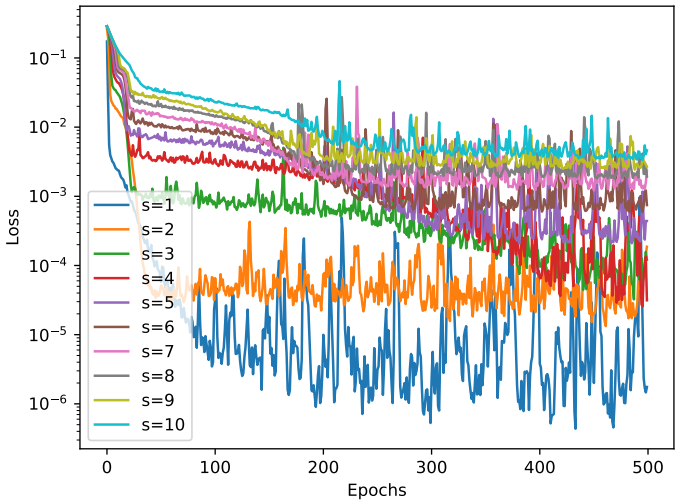
Second, we take a look at the losses during training when trying to approximate the remainder. These losses are plotted in fig. 5. In the cases (a) and (c), where the radius $R = 1$, we see almost a complete reversal of the order s. Previously, we had that the higher powers of s were higher but now the this trend has largely reversed. This has to do with what is being estimated. When $R = 1$ it does not take many Taylor terms to go to approximately zero. This means that whilst the network with $s = 1$ still has to fit f_g, the network for $s = 10$ has to approximate something that is already almost zero. In both (a) and (c) we see that there are different loss levels at which the neural networks settle. The higher of these two levels is the level of the final loss value for $s = 1$. This shows that the networks with higher order ReLU s small can capture the higher order terms of f_g roughly as well as the network with $s = 1$, and the networks with higher order ReLU s large can capture the higher order terms better than that. Since these networks with s large struggled in (a) and (c) of fig. 3 to approximate f_g, one could deduce from this that the networks with higher order ReLU s struggle to capture the polynomial part with order less than s. A new experiment should be conducted to investigate whether this is indeed true. In the cases (b) and (d), where $R = 3$, we again see roughly two levels to which the losses converge. The losses for $s = 1$ and $s = 2$ are very similar to what they are in (b) and (d) of fig. 3, whereas all the other losses are more concentrated on a single loss level than before. Similar to before, we can attribute the lower final losses to the neural networks not having to approximate the lower order polynomial terms of the Taylor expansion of f_g, and the higher losses in the $R = 3$ case compared to the $R = 1$ case to the number of Taylor terms required for the remainder to vanish on the interval.

Third and last, we plot eq. (326) and eq. (327) for the networks we have trained. These are plotted in fig. 6 and fig. 7 respectively. In these plots two things stand out. The first is that the actual loss values are orders of magnitude lower than the corresponding bounds from eq. (326) and eq. (327). This shows that the bound we have derived is not a tight bound. The second is that there is no real trend when we increase m or n, even though we expect that the bound decreases inversely with m and inversely with the square root of n from corollary 4.4.1. This suggests that the chosen m and n are large enough that the trained networks with the same s approximate f_g or the remainder $\tilde{f}_s$ with a similar error. Finally, note that the computed errors are above or around the set absolute error tolerance for 'scipy.quad'. This shows that the error to numerical integration indeed did not change the results.

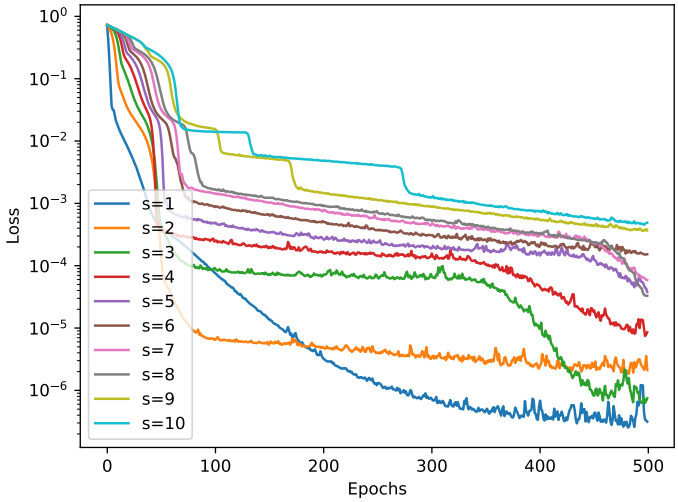
In summary, the bound in eq. (324) suggests that either you need a higher order ReLU with high enough order s or the ReLU itself to get the best results. The bound suggests that the order s for which the higher order ReLU outperforms the ReLU is dependent on the radius R. The previous simulations show that this does not happen in practice. Although this is just a single example and an even example with $d = 1$, it strongly suggests that ReLU will outperform the higher ReLU even when functions are smooth.



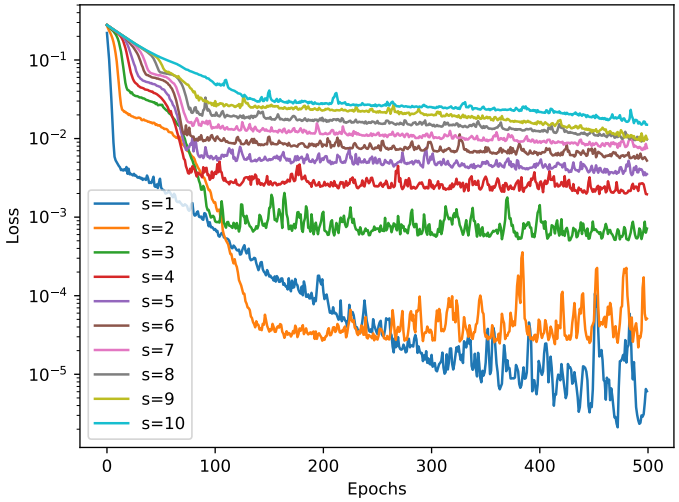
(a) Underparametrized with radius 1



(b) Underparametrized with radius 3

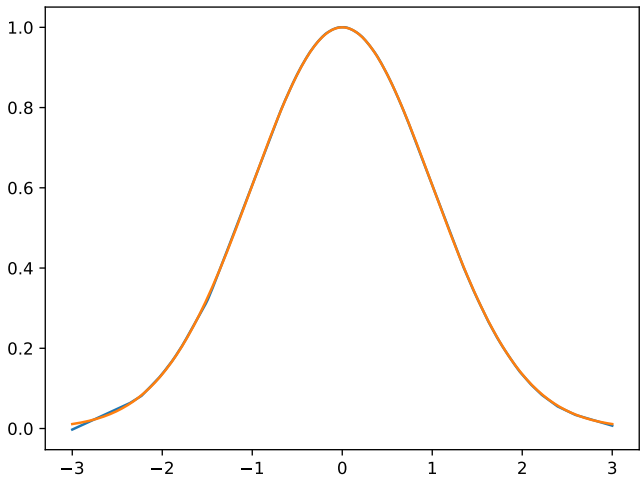


(c) Overparametrized with radius 1

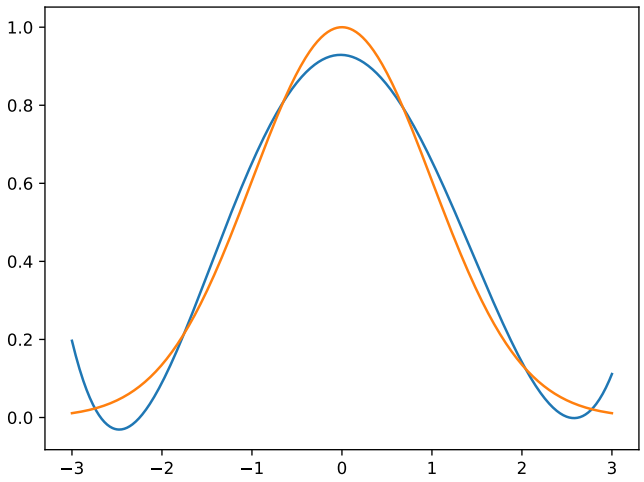


(d) Overparametrized with radius 3

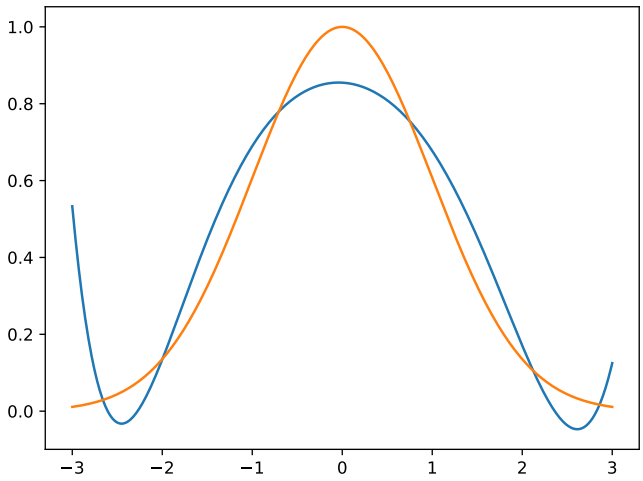
Figure 3: Training losses during training of f_g using a shallow neural network with $m = 500$ hidden neurons, higher order ReLU of order s as activation function, and in the under- and overparametrized regimes $n = 200$ and $n = 800$ respectively. On the left this is done with radius $R = 1$, and on the right with radius $R = 3$.



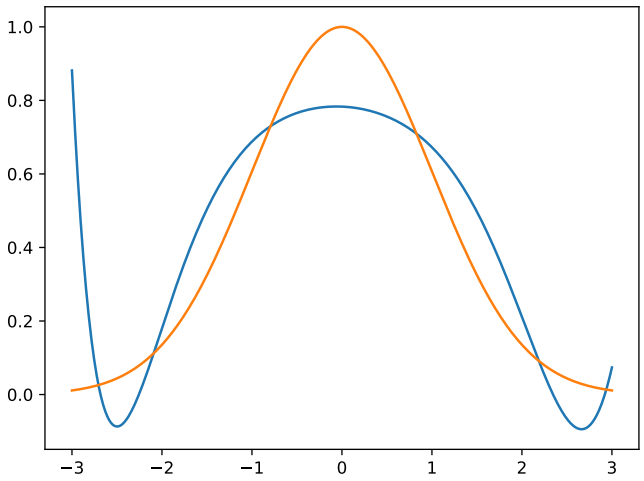
(a) Shallow neural network with higher order ReLU $s = 1$



(b) Shallow neural network with higher order ReLU $s = 4$

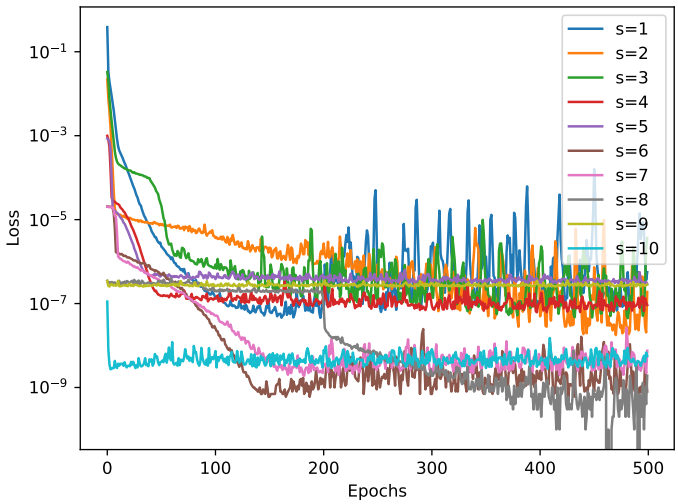


(c) Shallow neural network with higher order ReLU $s = 7$

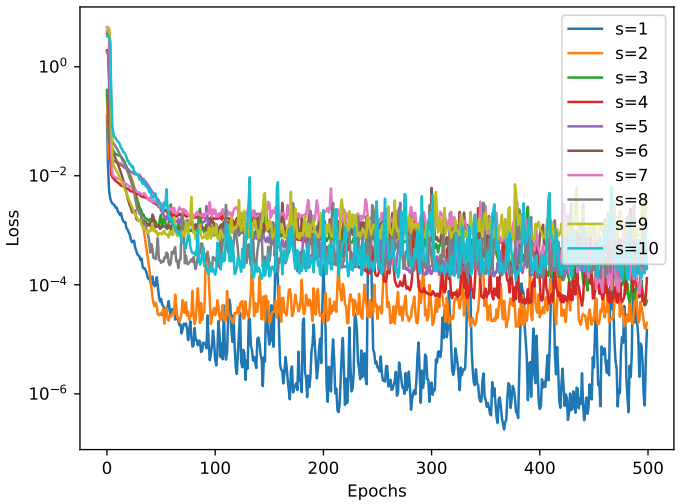


(d) Shallow neural network with higher order ReLU $s = 10$

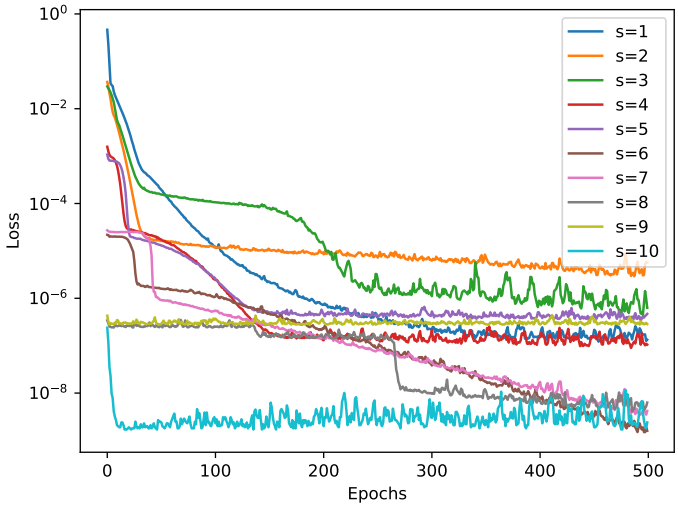
Figure 4: Shallow neural networks from fig. 3 (b) compared with f_g. The neural networks are indicated with blue, and f_g with orange.



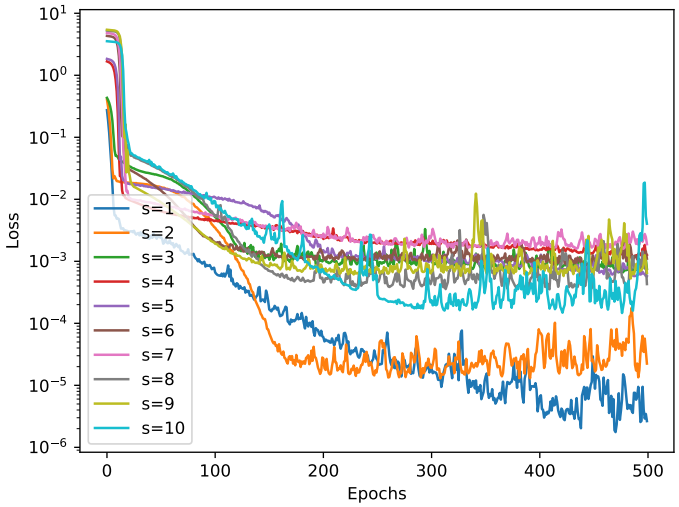
(a) Underparametrized with radius 1



(b) Underparametrized with radius 3

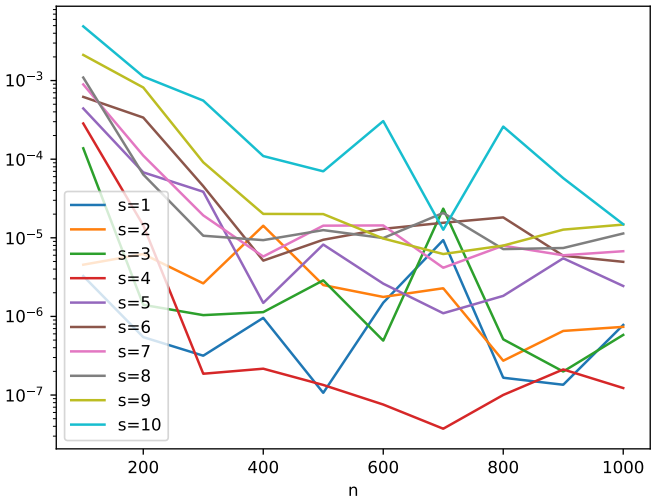


(c) Overparametrized with radius 1

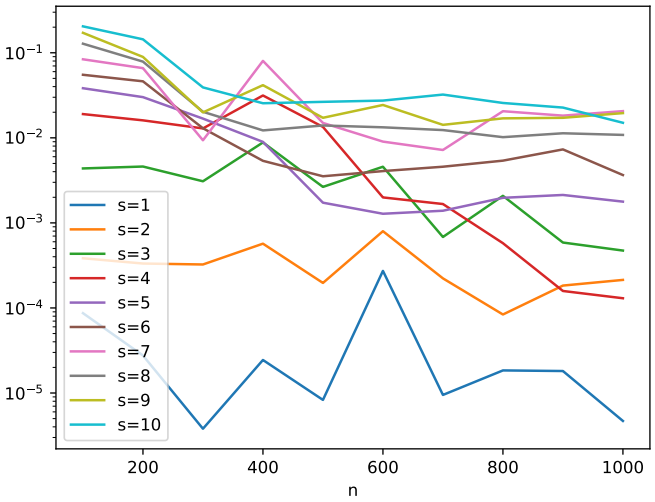


(d) Overparametrized with radius 3

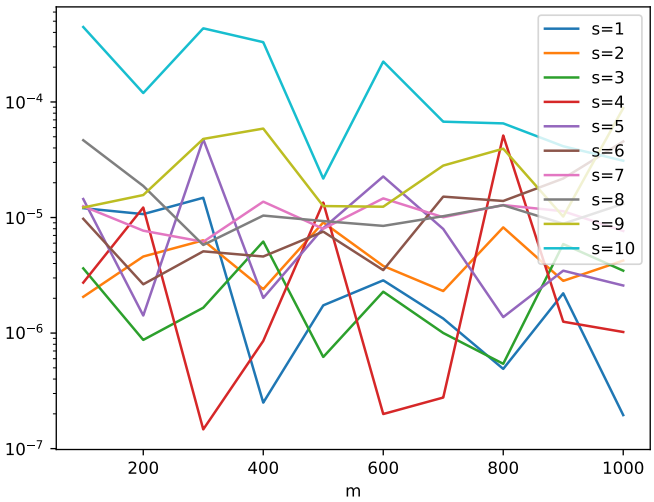
Figure 5: Training losses during training of the remainder $\tilde{f}_s$ using a shallow neural network with $m = 500$ hidden neurons, higher order ReLU of order s as activation function, and in the under- and overparametrized regimes $n = 200$ and $n = 800$ respectively. On the left this is done with radius $R = 1$, and on the right with radius $R = 3$.



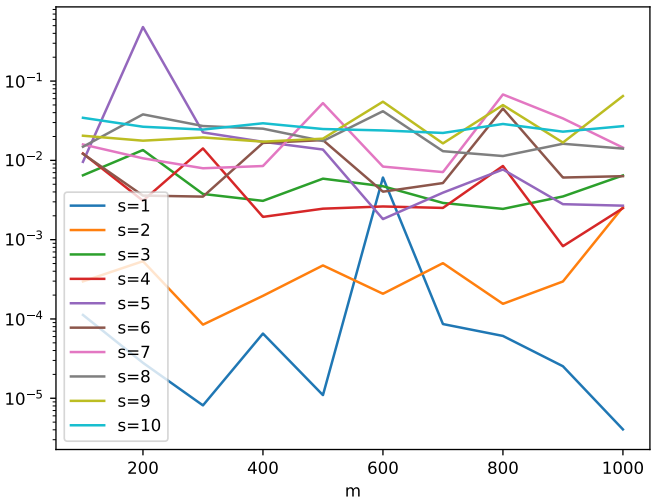
(a) L^2 norm with $m =$ fixed with radius $R = 1$



(b) L^2 norm with $m =$ fixed with radius $R = 3$

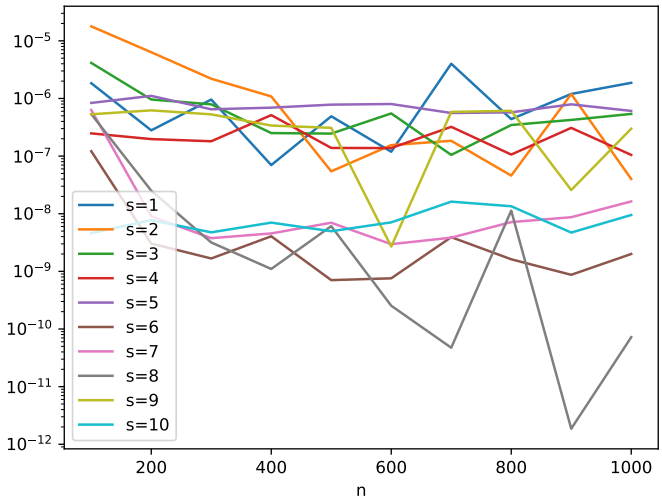


(c) L^2 norm with $n =$ fixed with radius $R = 1$

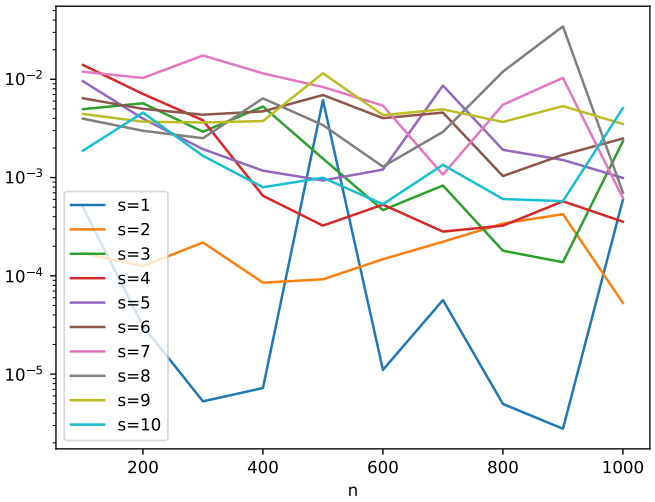


(d) L^2 norm with $n =$ fixed with radius $R = 3$

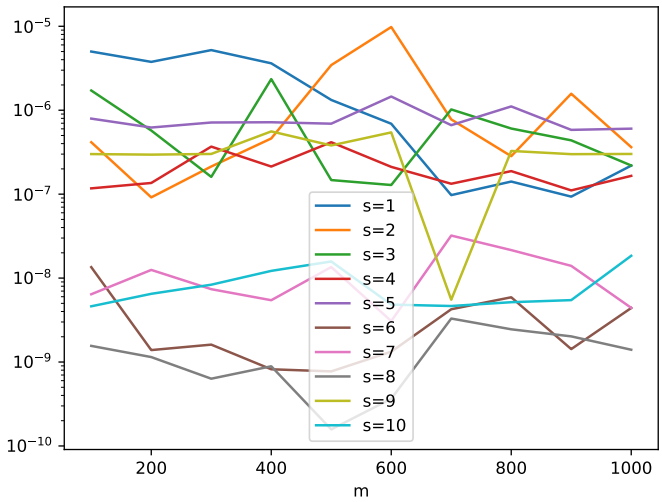
Figure 6: The L^2 norm of eq. (326) with a fixed number of neurons m with a varying number of samples n on top, and a varying number of neurons m with a fixed number of samples n below. The radius used is $R = 1$ on the left and $R = 3$ on the right. The L^2 norm is plotted against $1/m$ when varying the number of neurons m, and it is plotted against $1/\sqrt{n}$ when varying the number of samples n. In both cases this is done in accordance with the quantitative behaviour of eq. (324).



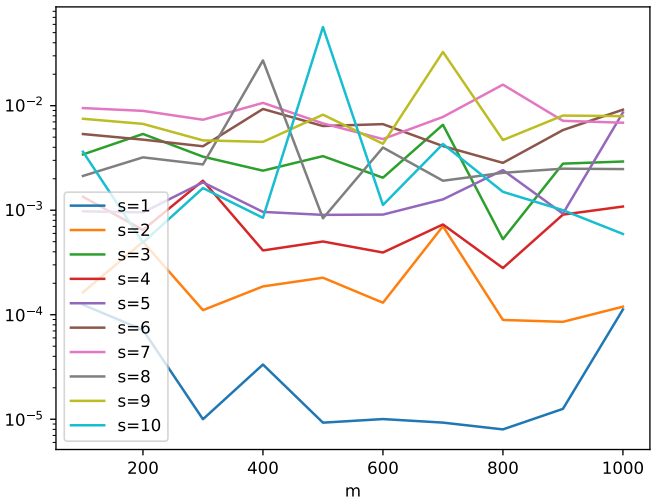
(a) L^2 norm with $m =$ fixed with radius $R = 1$



(b) L^2 norm with $m =$ fixed with radius $R = 3$



(c) L^2 norm with $n =$ fixed with radius $R = 1$



(d) L^2 norm with $n =$ fixed with radius $R = 3$

Figure 7: The L^2 norm of eq. (327) with a fixed number of neurons m with a varying number of samples n on top, and a varying number of neurons m with a fixed number of samples n below. The radius used is $R = 1$ on the left and $R = 3$ on the right. The L^2 norm is plotted against $1/m$ when varying the number of neurons m, and it is plotted against $1/\sqrt{n}$ when varying the number of samples n. In both cases this is done in accordance with the quantitative behaviour of eq. (324).

6 The Big Picture

In section 3 of this work we discussed relations between Barron spaces. In section 4 we introduced the Fourier based space $\mathscr{F}_I^{s,1}$. In this section we will take the various embeddings and inclusions for and between these spaces, and represent them in a couple of figures.

6.1 Bach and Barron

We will start with the embeddings and inclusions for and between Barron spaces. These are represented in fig. 8. We have discussed part of these relations in section 3, and another part of the relations has been proven by other authors. We will now give a proof for the remaining relations.

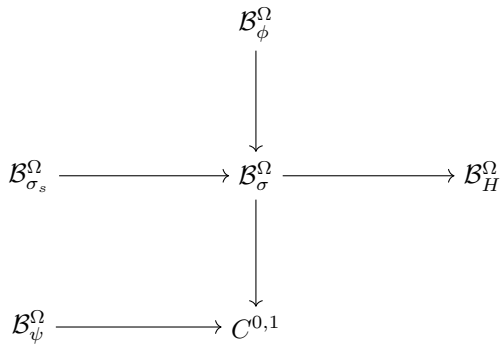


Figure 8: Relation between the various Fourier based spaces for $s, p \in \mathbb{N}$ with $s \geqslant p$, and $\phi \in C^2$. Arrows represent inclusions. The relation between $\mathcal{B}_{\sigma_s}$ and $\mathcal{B}_\sigma$ follows from theorem 3.2, the relation between $\mathcal{B}_\phi$ and $\mathcal{B}_\sigma$ from theorem 3.1, the relation between $\mathcal{B}_\sigma$ and $\mathcal{B}_H$ from [Caragea et al., 2020; lemma 7.1], and the relation between $\mathcal{B}_\phi$ and $C^{0,1}$ as well as the relation between $\mathcal{B}_\psi$ and $C^{0,1}$ from proposition 6.1.

We will first prove that Barron functions are Lipschitz. For the ReLU this was already shown in [E and Wojtowytsch, 2020b; theorem 3.3]. For the other activation functions this was not shown in other works, but from the proof that we will give now it is clear that the Barron norm was chosen such that Barron functions are guaranteed to be Lipschitz.

Proposition 6.1. *Let ϕ be an L-Lipschitz activation function and $f \in \mathcal{B}_\phi^\Omega$, then f is $L\|f\|_{\mathcal{B}_\phi^\Omega}$-Lipschitz.*

Proof. Let $x, y \in \mathcal{X}$ and $\mu \in M_{\phi,f}$, then

$$\begin{aligned} |f(x) - f(y)| &= \left| \int_\Omega \phi(\langle x|w\rangle + b) - \phi(\langle y|w\rangle + b) d\mu(w,b) \right| \\ &\leqslant \int_\Omega |\phi(\langle x|w\rangle + b) - \phi(\langle y|w\rangle + b)| d\mu(w,b) \end{aligned}$$

$$\begin{aligned}
&\leqslant \int_\Omega L|(\langle x|w\rangle + b) - (\langle y|w\rangle + b)|d\mu(w,b) \\
&= \int_\Omega L|\langle x-y|w\rangle|d\mu(w,b) \\
&\leqslant L\|x-y\| \int_\Omega \|w\|d\mu(w,b) \\
&\leqslant L\|x-y\| \int_\Omega \|w\| + |b|d\mu(w,b).
\end{aligned}$$

Taking the infimum over $\mu \in M_{\phi,f}$ gives

$$|f(x) - f(y)| \leqslant L\|f\|_{\mathcal{B}_\phi}\|x-y\|. \tag{328}$$

Q.E.D.

It might be tempting to add the inclusion of $\mathscr{F}_I^{2,1}$ into $\mathcal{B}_\sigma^\Omega$ to fig. 8. After all, we have shown in section 4.3 that on balls of finite radius R we can find, for each function $f \in \mathscr{F}_I^{2,1}$, a measure $\mu \in rca(\Omega)$ such that

$$f(x) = f(0) + \sum_{|\alpha|=1} \partial^\alpha f(0)x^\alpha + \int_{\mathbb{S}^{d+1}\times[0,R]} \sigma(\langle x|w\rangle + b)d\mu(w,b) \tag{329}$$

with α a multi-index and μ satisfying

$$\|\mu\|_{W_\sigma,\mathbb{S}^{d+1}\times[0,R]} \leqslant 2(1+R)\|f\|_{\mathscr{F}^{2,1}}. \tag{330}$$

Since $|\alpha| = 1$, at most 1 component of α will be 1 and the rest will be 0. That means that $x^\alpha = x_i$ for some i depending on α. Similarly, $\partial^\alpha f(0) = \partial_i f(0)$. Hence, when we define the measures

$$\mu_0 = f(0)\delta_{0,1} \tag{331}$$

and

$$\mu_i = \partial_i f(0)\delta_{e_i,0} \tag{332}$$

for $i \in \{1,...,d\}$ with $e_i \in \mathbb{R}^d$ a vector with only a 1 on the ith component and zeroes elsewhere, we can combine μ_0, μ_i and μ into a single measure

$$\nu = \mu_0 + \sum_{i=1}^d \mu_i + \mu \tag{333}$$

with

$$\|\nu\|_{W_\sigma,\mathbb{S}^{d+1}\times[0,R]} < \infty \tag{334}$$

such that

$$f = K_\sigma^\Omega \nu \tag{335}$$

on that specific ball of radius R. Unfortunately, functions $f \in \mathscr{F}_I^{2,1}$ do not have to be compactly supported, so we cannot choose R big enough and construct ν like above. It might be possible to construct a localised version of $\mathscr{F}_I^{2,1}$ such that this does work. This will be further discussed in section 8.

6.2 Fourier based spaces

We have talked a fair bit about the Fourier based space $\mathscr{F}_I^{s,1}$. In this section we will discuss the smoothness properties of that space. In particular, we will look at inclusions and embeddings with respect to the Sobolev spaces $W^{p,k}$ and $\mathcal{H}^p$ and the continuous differentiable functions C^s. These inclusions and embeddings are represented in fig. 9. We will now prove the relations that have not been proven before, and end with a discussion about the Fourier space version of the Sobolev spaces and its relation to $\mathscr{F}_I^{s,1}$.

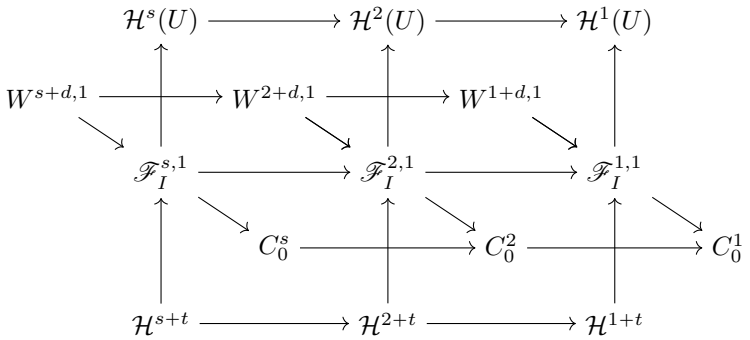


Figure 9: Relation between the Fourier based spaces $\mathscr{F}_I^{s,1}$, the continuous spaces C^s and the Sobolev spaces H^s, H^{s+t} and $W^{s+d,1}$ for $s, t \in \mathbb{N}$ with $t > d/2$. All spaces are over $\mathbb{R}^d$, except for the Sobolev spaces $\mathcal{H}^s(U)$, where U is a compact subset of $\mathbb{R}^d$. Arrows represent inclusions. The inclusions of Sobolev spaces into other Sobolev spaces follow directly from their definition. The same holds for the continuously differentiable functions and $\mathscr{F}_I^{s,1}$ itself. The relation between $\mathscr{F}_I^{s,1}$ and H^{s+t} follows from proposition 6.2, the relation between $\mathscr{F}_I^{s,1}$ and $W^{s+d,1}$ from proposition 6.3, the relation between $\mathscr{F}_I^{s,1}$ and C_0^s from proposition 6.4, and the relation between $\mathscr{F}_I^{s,1}$ and $W^{s+d,1}$ from proposition 6.5

.

We will start with the inclusions and embeddings into $\mathscr{F}_I^{s,1}$. These are given in the following two propositions. The first shows us the number of L^2 weak derivatives that is sufficient for a function to be in $\mathscr{F}_I^{s,1}$.

Proposition 6.2. *Let $s \in \mathbb{N}$ and $t > d/2$, then*

$$H^{s+t}(\mathbb{R}^d) \hookrightarrow \mathscr{F}_I^{s,1}. \tag{336}$$

Proof. Let $f \in H^{s+t}$. We will show that

$$\int_{\mathbb{R}^d} (1 + \|\xi\|^s)|\hat{f}(\xi)|d\xi < \infty. \tag{337}$$

This is sufficient since

$$1 + \|\xi\|^s \leqslant (1 + \|\xi\|)^s \leqslant 2^s(1 + \|\xi\|^s) \tag{338}$$

implies that $\|f\|_{\mathscr{F}_I^{1,1}}$ is equal to eq. (337) up to a constant.

By Cauchy Schwartz and Plancherel's theorem we obtain

$$\begin{aligned}\int_{\mathbb{R}^d}(1+\|\xi\|^s)|\hat{f}(\xi)|d\xi &= \int_{\mathbb{R}^d}(1+\|\xi\|^s)|\hat{f}(\xi)|(1+\|\xi\|^{2t})^{1/2}(1+\|\xi\|^{2t})^{-1/2}d\xi \\ &\leqslant \left(\int_{\mathbb{R}^d}(1+\|\xi\|^s)^2|\hat{f}(\xi)|^2(1+\|\xi\|^{2t})d\xi\right)\left(\int_{\mathbb{R}^d}\frac{1}{1+\|\xi\|^{2t}}d\xi\right) \\ &= \left(\int_{\mathbb{R}^d}(1+2\|\xi\|^s+\|\xi\|^{2s})|\hat{f}(\xi)|^2(1+\|\xi\|^{2t})d\xi\right)\left(\int_{\mathbb{R}^d}\frac{1}{1+\|\xi\|^{2t}}d\xi\right) \\ &= \left(\int_{\mathbb{R}^d}(1+2\|\xi\|^s+\|\xi\|^{2s})|\hat{f}(\xi)|^2(1+\|\xi\|^{2t})d\xi\right)\left(\int_{\mathbb{R}^d}\frac{1}{1+\|\xi\|^{2t}}d\xi\right) \\ &\leqslant 2\left(\int_{\mathbb{R}^d}|\hat{f}(\xi)|^2\sum_{k=0}^{s+t}\|\xi\|^{2k}d\xi\right)\left(\int_{\mathbb{R}^d}\frac{1}{1+\|\xi\|^{2t}}d\xi\right).\end{aligned}$$

Since $f\in H^{s+t}$, the first factor is finite. The second factor satisfies

$$\int_{\mathbb{R}^d}\frac{1}{1+\|\xi\|^{2t}}d\xi = C\int_0^\infty\frac{r^{d-1}}{1+r^{2t}}dr<\infty \tag{339}$$

for some $0<C<\infty$ by the condition on t. Therefore, $f\in\mathscr{F}^{s,1}$ and

$$\|f\|_{\mathscr{F}^{s,1}}\leqslant 2^{s+1}\left(\int_{\mathbb{R}^d}\frac{1}{1+\|\xi\|^{2t}}d\xi\right)\|f\|_{H^{s+t}}. \tag{340}$$

. *Q.E.D.*

The second shows us the number of L^1 weak derivatives that is sufficient for a function to be in $\mathscr{F}_I^{s,1}$.

Proposition 6.3. *Let $s\in\mathbb{N}$, then*

$$W^{s+d,1}(\mathbb{R}^d)\hookrightarrow\mathscr{F}_I^{s,1}. \tag{341}$$

Proof. This proof relies on [Kolyada, 1997; theorem A]. It states that the inequality

$$\int_{\mathbb{R}^d}\|\xi\|^{r-d}\left|\hat{f}(\xi)\right|d\xi\leqslant\sum_{|\alpha|=r}\|D^\alpha f\|_{L^1(\mathbb{R}^d)} \tag{342}$$

holds for $f\in W^{r,1}$, where α is an multi-index. Hence, for $f\in W^{r+d,1}$ we have

$$\int_{\mathbb{R}^d}\|\xi\|^{r}\left|\hat{f}(\xi)\right|d\xi\leqslant\sum_{|\alpha|=r+d}\|D^\alpha f\|_{L^1(\mathbb{R}^d)}. \tag{343}$$

Since $W^{s+d,1}\subseteq W^{r+d,1}$ for all $0\leqslant r\leqslant s$,

$$\|f\|_{\mathscr{F}^{s,1}}\leqslant 2^s\int_{\mathbb{R}^d}(1+\|\xi\|^s)\left|\hat{f}(\xi)\right|d\xi\leqslant 2^s\|f\|_{W^{s+d,1}(\mathbb{R}^d)} \tag{344}$$

for $f\in W^{s+d,1}$. *Q.E.D.*

Next up are the inclusions and embeddings of $\mathscr{F}_I^{s,1}$ into other spaces. These are given in the following two propositions. The first tells us about the minimal smoothness $f \in \mathscr{F}_I^{s,1}$ must have.

Proposition 6.4. *Let $s \in \mathbb{N}$, then*

$$\mathscr{F}_I^{s,1} \hookrightarrow C_0^s(\mathbb{R}^d). \tag{345}$$

Proof. Let $|\alpha| \leqslant s$, and $f \in \mathscr{F}_I^{s,1}$. With this we have that $\xi \mapsto \xi^\alpha \hat{f}(\xi) \in L^1$, because

$$\begin{aligned} \int_{\mathbb{R}^d} \left|\xi^\alpha \hat{f}(\xi)\right| d\xi &\leqslant \int_{\mathbb{R}^d} |\xi|^{|\alpha|} \left|\hat{f}(\xi)\right| d\xi \\ &= \|f\|_{\mathscr{F}^{|\alpha|,1}} \\ &\leqslant \|f\|_{\mathscr{F}_I^{s,1}} < \infty. \end{aligned} \tag{346}$$

If $x_k \in \mathbb{R}^d$ is a sequence such that $x_k \to x \in \mathbb{R}^d$, then by continuity of $x \mapsto e^x$

$$\lim_{k\to\infty} \left| e^{i\langle x|\xi_k\rangle} - e^{i\langle x|\xi\rangle} \right| = 0 \tag{347}$$

for all $\xi \in \mathbb{R}^d$. At the same time

$$\left| e^{i\langle x|\xi_k\rangle} - e^{i\langle x|\xi\rangle} \right| \leqslant 2 \tag{348}$$

for all $\xi \in \mathbb{R}^d$. Hence, by the dominated convergence theorem

$$\lim_{k\to\infty} \left| \int_{\mathbb{R}^d} \xi^\alpha \hat{f}(\xi) e^{i\langle x_k|\xi\rangle} d\xi - \int_{\mathbb{R}^d} \xi^\alpha \hat{f}(\xi) e^{i\langle x|\xi\rangle} d\xi \right| \leqslant \lim_{k\to\infty} \int_{\mathbb{R}^d} \left|\xi^\alpha \hat{f}(\xi)\right| \left| e^{i\langle x_k|\xi\rangle} - e^{i\langle x|\xi\rangle} \right| d\xi = 0 \tag{349}$$

This shows that $x \mapsto \int_{\mathbb{R}^d} \xi^\alpha \hat{f}(\xi) e^{i\langle x|\xi\rangle} d\xi$ is continuous. To show that f decays properly, we use that the compactly supported smooth functions $C_c^\infty(\mathbb{R}^d)$ are dense in $W^{k,p}(\mathbb{R}^d)$ for all $k \geqslant 0$ and $1 \leqslant p < \infty$ [Tao, 2009; lemma 23]. For fixed $x \in \mathbb{R}^d$ with $\|x\|_{\ell^\infty} \geqslant R > 0$ there must be a component x_j of x such that $x_j \geqslant R$. Hence, for all $\hat{g} \in C_c^\infty(\mathbb{R}^d)$ we have that

$$\begin{aligned} \left| \int_{\mathbb{R}^d} \hat{g}(\xi) e^{i\langle x|\xi\rangle} d\xi \right| &= \left| \int_{\mathbb{R}^d} \partial_j(\hat{g}(\xi)) \frac{1}{ix_j} e^{i\langle x|\xi\rangle} d\xi \right| \\ &\leqslant \int_{\mathbb{R}^d} |\partial_j(\hat{g}(\xi))| \left| \frac{1}{ix_j} \right| \left| e^{i\langle x|\xi\rangle} \right| d\xi \\ &\leqslant \frac{1}{R} \int_{\mathbb{R}^d} |\partial_j(\hat{g}(\xi))| d\xi \\ &= \frac{\|\partial_j(\hat{g}(\xi))\|_{L^1(\mathbb{R}^d)}}{R} < \infty. \end{aligned} \tag{350}$$

Taking the limit of R to infinity gives

$$\lim_{R\to\infty} \left| \int_{\mathbb{R}^d} \hat{g}(\xi) e^{i\langle x|\xi\rangle} d\xi \right| = 0. \tag{351}$$

Since $\hat{g} \in C_c^\infty(\mathbb{R}^d)$ is arbitrary and $C_c^\infty(\mathbb{R}^d)$ is dense in $L^1(\mathbb{R}^d)$, it must also hold that $x \mapsto \int_{\mathbb{R}^d} \xi^\alpha \hat{f}(\xi) e^{i\langle x|\xi\rangle} d\xi$ vanishes at infinity.

$x \mapsto \int_{\mathbb{R}^d} \xi^\alpha \hat{f}(\xi) e^{i\langle x|\xi\rangle} d\xi$ is continuous and vanishes at infinity and $\xi \mapsto \xi^\alpha \hat{f}(\xi) \in L^1$, so we have

$$\partial^\alpha f(x) = \partial^\alpha \int_{\mathbb{R}^d} \hat{f}(\xi) e^{i\langle x|\xi\rangle} d\xi = \int_{\mathbb{R}^d} \xi^\alpha \hat{f}(\xi) e^{i\langle x|\xi\rangle} d\xi. \tag{352}$$

for all $x \in \mathbb{R}^d$. This implies that $f \in C_0^s(\mathbb{R}^d)$ and that

$$\begin{aligned} |\partial^\alpha f(x)| &= \left| \partial^\alpha \int_{\mathbb{R}^d} \hat{f}(\xi) e^{i\langle x|\xi\rangle} d\xi \right| \\ &= \left| \int_{\mathbb{R}^d} \xi^\alpha \hat{f}(\xi) e^{i\langle x|\xi\rangle} d\xi \right| \\ &\leqslant \int_{\mathbb{R}^d} \left| \xi^\alpha \hat{f}(\xi) \right| d\xi \qquad \textit{eq.}\ (349) \\ &\leqslant \|f\|_{\mathscr{F}_I^{s,1}}. \end{aligned} \tag{353}$$

for all $x \in \mathbb{R}^d$. Since x and α are arbitrary,

$$\|f\|_{C^s(\mathbb{R}^d)} \leqslant \|f\|_{\mathscr{F}_I^{s,1}}.$$

Q.E.D.

The combinations of proposition 6.2 and proposition 6.3 with proposition 6.4 give

$$H^{s+t}(\mathbb{R}^d) \hookrightarrow \mathscr{F}_I^{s,1} \hookrightarrow C_0^s(\mathbb{R}^d) \tag{354}$$

and

$$W^{s+d,1}(\mathbb{R}^d) \hookrightarrow \mathscr{F}_I^{s,1} \hookrightarrow C_0^s(\mathbb{R}^d) \tag{355}$$

respectively. On the other hand, the Sobolev embedding states that

$$W^{s+d,1}(\mathbb{R}^d) \hookrightarrow C_0^s(\mathbb{R}^d) \tag{356}$$

and

$$H^{s+t}(\mathbb{R}^d) = W^{s+t,2}(\mathbb{R}^d) \hookrightarrow C_0^{s,t-d/2}(\mathbb{R}^d) \hookrightarrow C_0^s(\mathbb{R}^d) \tag{357}$$

when $t - d/2 \in [0, 1]$ ["Sobolev inequality", 2021; Visintin, 2017]. Hence, the combinations of proposition 6.3 and proposition 6.2 with proposition 6.4 can be seen as a proof for the special cases of the Sobolev embeddings theorem on $\mathbb{R}^d$ using $\mathscr{F}_I^{s,1}$ as an intermediate space.

The second shows us the number of L^2 weak derivatives that is necessary for a function to be in $\mathscr{F}_I^{s,1}$. Proposition 6.5 is an adaptation of lemma 2 and the surrounding text from [Siegel and Xu, 2021]. The notation has been changed to align with this work, and has been altered to prove an embedding instead of just an upper bound. This proof relies on the Schwartz functions $S(\mathbb{R}^d)$. These are given by

$$\begin{aligned} S(\mathbb{R}^d) &= \left\{ f \in C^\infty(\mathbb{R}^d, \mathbb{C}) \;\middle|\; \forall \alpha, \beta \in \mathbb{N}^d :\ \|f\|_{\alpha,\beta} < \infty \right\}, \\ \|f\|_{\alpha,\beta} &= \sup_{x \in \mathbb{R}^d} \left| x^\alpha \partial^\beta f(x) \right|. \end{aligned} \tag{358}$$

A Schwartz function $f \in S(\mathbb{R}^d)$ can be thought of as a function for which all of its derivatives vanish faster than any reciprocal power of x.

Proposition 6.5 (Adaptation of [Siegel and Xu, 2021; lemma 2]). *Let $s \in \mathbb{N}$ and be $U \subset \mathbb{R}^d$ a compact set, then $\mathscr{F}_I^{s,1} \hookrightarrow H^s(U)$.*

Proof. The norm of $\mathscr{F}_I^{s,1}$ is a polynomial weighted L^1 norm in Fourier space. The Schwartz functions are dense in this space. At the same time, Schwartz functions are also in $\mathcal{H}^s(U)$; This means that it is sufficient to proof that

$$\|f\|_{H^s(\mathcal{X})} \leqslant C\|f\|_{\mathscr{F}_I^{s,1}} \tag{359}$$

for all Schwartz functions.

Let $\chi_{\mathcal{X}}$ be the characteristic function of $\mathcal{X}$ and α be a multi-index with $|\alpha| \leqslant s$, then

$$\|D^\alpha f\|_{L^2(\mathcal{X})} = \|\chi_{\mathcal{X}} D^\alpha f\|_{L^2(\mathbb{R}^d)} = \left\|\hat{\chi}_{\mathcal{X}} * \widehat{D^\alpha f}\right\|_{L^2(\mathbb{R}^d)}. \tag{360}$$

An application of Young's inequality for convolutions and Plancherel's theorem gives

$$\left\|\hat{\chi}_{\mathcal{X}} * \widehat{D^\alpha f}\right\|_{L^2(\mathbb{R}^d)} \leqslant \|\hat{\chi}_{\mathcal{X}}\|_{L^2(\mathbb{R}^d)} \left\|\widehat{D^\alpha f}\right\|_{L^1(\mathbb{R}^d)} = \|\chi_{\mathcal{X}}\|_{L^2(\mathbb{R}^d)} \left\|\widehat{D^\alpha f}\right\|_{L^1(\mathbb{R}^d)} = |\mathcal{X}|^{\frac{1}{2}} \left\|\widehat{D^\alpha f}\right\|_{L^1(\mathbb{R}^d)}. \tag{361}$$

Using the Fourier differentiation identity and eq. (346) we see that the right most term satisfies

$$\left\|\widehat{D^\alpha f}\right\|_{L^1(\mathbb{R}^d)} = \left\|\xi^\alpha \hat{f}\right\|_{L^1(\mathbb{R}^d)} \leqslant \|f\|_{\mathscr{F}_I^{s,1}}. \tag{362}$$

Hence, after summing over all the multi-indices α with $|\alpha| \leqslant s$

$$\|f\|_{H^s(\mathcal{X})} = \sum_{|\alpha|=0}^{s} \|D^\alpha f\|_{L^2(\mathcal{X})} \leqslant \sum_{|\alpha|=0}^{s} |\mathcal{X}|^{\frac{1}{2}} \|f\|_{\mathscr{F}_I^{s,1}} = C\|f\|_{\mathscr{F}_I^{s,1}} \tag{363}$$

for some finite $C > 0$. *Q.E.D.*

We have now proven all the relations of fig. 9. We will finish this section with an observation between the Fourier version of Sobolev spaces and $\mathscr{F}_I^{s,1}$. The Fourier version of Sobolev spaces are given by

$$\begin{aligned} \hat{W}^{s,p}(\mathbb{R}^d) &= \left\{ u \in S^*(\mathbb{R}^d) \;\middle|\; \|u\|_{\hat{W}^{s,p}(\mathbb{R}^d)} < \infty \right\} \\ \|u\|_{\hat{W}^{s,p}(\mathbb{R}^d)} &= \left\| \left((1 + \|\cdot\|^2)^{s/2} \hat{u} \right)^\vee \right\|_{L^p(\mathbb{R}^d)} \end{aligned} \tag{364}$$

where $\vee$ indicates the inverse Fourier transform and $S^*(\mathbb{R}^d)$ is the space of tempered distributions, the dual space of the Schwartz functions $S(\mathbb{R}^d)$. For $s \in \mathbb{N}$ and $1 < p < \infty$ we have that

$$\hat{W}^{s,p}(\mathbb{R}^d) \cong W^{s,p}(\mathbb{R}^d), \tag{365}$$

see [Grafakos, 2014b; section 1.3.1] for more information. When we take $p = 2$, then by Plancherel's theorem

$$\|u\|_{\hat{W}^{s,2}(\mathbb{R}^d)} = \left\| \left((1 + \|\cdot\|^2)^{s/2} \hat{u} \right)^\vee \right\|_{L^2(\mathbb{R}^d)} = \left\| (1 + \|\cdot\|^2)^{s/2} \hat{u} \right\|_{L^2(\mathbb{R}^d)}. \tag{366}$$

Since

$$(1 + \|\cdot\|^2)^{s/2} \leqslant 2^{s/2}(1 + \|\xi\|)^s \leqslant 2^{3s/2}(1 + \|\cdot\|^2)^{s/2}, \tag{367}$$

eq. (366) can be seen as an L^2 version of $\|f\|_{\mathscr{F}_I^{s,1}}$. Sadly, it is not clear how we could learn more about $\mathscr{F}_I^{s,1}$ from this due to two things. Firstly,

$$\hat{W}^{s,1}(\mathbb{R}^d) \ncong W^{s,1}(\mathbb{R}^d), \tag{368}$$

since $\xi \mapsto (1 + \|\xi\|^2)^{s/2}$ is not a Fourier multiplier for $p = 1$ [Grafakos, 2014a; section 6.2.3 and in particular theorem 6.2.7]. This means that if we can establish a link from $\mathscr{F}_I^{s,1}$ to $\hat{W}^{s,p}(\mathbb{R}^d)$, it does not tell us much about the more interesting space $W^{s,p}(\mathbb{R}^d)$. Secondly, Plancherel's theorem requires $p = 2$. This means that for $p \neq 2$, we have an additional inverse Fourier transform in the norm. The Hausdorff-Young inequality can be used to get rid of this inverse Fourier transform. It states that

$$\left\|\hat{f}\right\|_{L^q(\mathbb{R}^d)} \leqslant \|f\|_{L^p(\mathbb{R}^d)} \tag{369}$$

for $\frac{1}{p} + \frac{1}{q} = 1$, $p \in [1, 2]$ and $f \in L^p(\mathbb{R}^d)$. In this form we cannot get $\mathscr{F}_I^{s,1}$ out, since $q \geqslant 2$ and we need $q = 1$. However, if we take $p = 1$ and $\hat{f} \in L^p(\mathbb{R}^d)$, then by symmetry we get

$$\left\|(\hat{f})^\vee\right\|_{L^\infty(\mathbb{R}^d)} = \left\|\hat{\hat{f}}\right\|_{L^\infty(\mathbb{R}^d)} \leqslant \left\|\hat{f}\right\|_{L^1(\mathbb{R}^d)}. \tag{370}$$

Replacing $\hat{f}$ with $(1 + \|\xi\|)^s \hat{f}$ and subsequently using eq. (367), allows us to write

$$\begin{aligned}
\|f\|_{W^{s,\infty}} &= \|f\|_{\hat{W}^{s,\infty}} \\
&= \left\|\left((1 + \|\cdot\|^2)^{s/2}\hat{f}\right)^\vee\right\|_{L^\infty(\mathbb{R}^d)} \\
&\leqslant 2^{s/2}\left\|\left((1 + \|\cdot\|)^s\hat{f}\right)^\vee\right\|_{L^\infty(\mathbb{R}^d)} \\
&\leqslant 2^{s/2}\left\|\widehat{(1 + \|\cdot\|)^s\hat{f}}\right\|_{L^\infty(\mathbb{R}^d)} \\
&\leqslant 2^{s/2}\left\|\hat{f}\right\|_{\mathscr{F}_I^{s,1}}.
\end{aligned}$$

This is already implied by proposition 6.4. Hence, although the Hausdorff-Young inequality gets rid of the inverse Fourier transform, it does not provide us with new information. It is unknown to the author whether there are other Fourier theorems that could get rid of the additional inverse Fourier transform in the norm.

7 Deep Learning and Control

In the previous sections we have taken a look at Bach and Barron spaces. These are infinitely wide shallow neural networks. However, in practice we don't use shallow neural networks; we use deep neural networks instead. These deep neural networks consists of several shallow networks, with proper input and output sizes, put back to back. This process makes it so that deep neural networks can approximate more functions than shallow neural networks. These deep neural networks are starting to be used in more and more fields of research. One of these research areas is that of control. In the following sections we will take a closer look at the interplay between deep learning and control. We will demonstrate that this interplay goes both ways.

7.1 Deep Learning in Control

In control the goal is to steer a physical system, typically called a *plant*, into a certain configuration or to follow a certain trajectory through space. This configuration or trajectory is called the *reference*. To achieve this a control signal is added to the system. The control is added, so that a certain cost functional that includes the reference, is minimised and, at the same time, no constraints are broken. These constraints include limitations on the size of the control signal as well as infeasible or undesirable configurations and trajectories. Each of the branches of control theory deals with one or more aspects of this minimisation problem. In optimal control the goal is to find an optimal solution, in robust control the goal is to find a good enough control signal that is still close to optimal and satisfies the constraints when there is some noise or disturbance, and in systems identification the goal is to find a model and parameters that properly approximate the behaviour of the plant. Solution strategies have been determined for many problems. At one point or another, many of these strategies involve approximating a function or model. Difficulties arise when these functions or models are highly non-linear or high dimensional. In these cases finding a proper approximation can become a time intensive and computationally expensive task. These difficult to approximate functions and models can be approximated efficiently using deep learning methods.

We will now take a specific problem to exemplify how this works in practice. We will describe the problem using a model-based method and a model-free method. We will solve the problem using these methods from a control perspective and identify the parts in which learning methods as a function approximation tool would be most useful. In particular, we consider the problem of flying a large number of drones autonomously from the ground to a particular configuration in the sky. The route they will fly will not be clear of obstacles and there will be wind, but they will be aware of the terrain. We want these drones to fly to their desired location without bumping into each other, without crashing and without taking an unnecessarily long route. This is a high dimensional and non-linear problem, which is actively being researched due to its highly complex nature and its high societal relevance. [Batra et al., 2021] includes videos that demonstrate what we want the drones to do visually. In section 7.1.1 we will discuss how to solve it using the model-based method called model adaptive control, and in section 7.1.2 we will discuss how to solve it using the model-free method called approximate dynamic programming. The main points of both of these sections are highlighted in section 7.1.3.

7.1.1 Model adaptive control

In model adaptive control the controller is designed based on a nominal model of the plant, and the controller is changed in each time step to handle previous disturbances as well as minimise the effect of predicted disturbances. To solve a problem it is typically broken into pieces. For the drone swarm it means that we consider the following pieces:

1. A single drone without friction,
2. a drone with drag,
3. a drone with ground effect,
4. several drones with downwash.

In section 7.1.1.1 till 7.1.1.4 we discuss these 4 cases. We assume that there is a suitable low level controller in place that ensures the rotors of the drone spin in such a way that the desired forces and torques of the controllers we will discuss is achieved. We will also assume reference trajectories are given, so we only have to design the controller.

7.1.1.1 Single drone – frictionless

If the drone is high enough above the ground and there is no air resistance aside from the resistance that allows the rotors to keep the drone up, then we can write a model for its equations of motion. Let

$$x_t = (p_t, v_t, R_t, \omega_t) \in \mathbb{R}^3 \times \mathbb{R}^3 \times SO(3) \times \mathbb{R}^3 \tag{371}$$

be the state vector with the position, velocity, orientation and angular velocities of the drone, then the equations of motion are given by

$$\begin{aligned} \partial_t p_t &= v_t & m\partial_t v_t &= mg + R_t u_t^f \\ \partial_t R_t &= R[\omega]_\times & J\partial_t \omega_t &= J\omega \times \omega + u_t^\tau \end{aligned} \tag{372}$$

where m is the mass, J is the inertia matrix, $[\cdot]_\times$ is the skew symmetric mapping of the cross product, $g = [0, 0, -g]^\intercal$ is the gravity vector, and u_t^f and u_t^τ are the forces and torques due to the rotors of the drone respectively [Shi et al., 2019]. The general robotics dynamics model is given by

$$M(q_t)\partial_t^2 q_t + C(q_t, \partial_t q_t)\partial_t q_t + G(q_t) - B(q_t)u_t = 0, \tag{373}$$

with q_t the generalised coordinates, M the inertial matrix, C the certifugal and Coriolis effects, G the gravitational terms, and B the attenuation matrix. Equation (372) can be written as eq. (373) by taking

$$\begin{aligned} q_t &= \begin{pmatrix} p_t & \theta_t \end{pmatrix}^\intercal \\ M(q_t) &= \begin{pmatrix} mI & 0 \\ 0 & J \end{pmatrix} \end{aligned}$$

$$C(q_t, \partial_t q_t) = \begin{pmatrix} 0 & 0 \\ 0 & J[\omega]_\times \end{pmatrix}$$
$$G(q_t) = \begin{pmatrix} mg & 0 \end{pmatrix}^\intercal$$
$$B(q_t) = \begin{pmatrix} R_t & 0 \\ 0 & I \end{pmatrix}$$

with I the identity matrix and θ_t representing the Euler angles. The suitable controller is then given by

$$u_t = B^\dagger \Big(M(q_t)\partial_t^2 p_t + C(q_t, \partial_t q_t)\partial_t q_t + G(q_t) \Big), \tag{374}$$

with $B^\dagger$ the Moore-Penrose pseudoinverse of B. If there is a disturbance $d(q_t, \partial_t q_t)$, then the general robotics dynamics model is given by

$$M(q_t)\partial_t^2 q_t + C(q_t, \partial_t q_t)\partial_t q_t + G(q_t) - B(q_t)u_t = d(q_t, \partial_t q_t). \tag{375}$$

If one has a good estimate $\hat{d}(q_t, \partial_t q_t)$ of the disturbance $d(q_t, \partial_t q_t)$, then a suitable controller is given by

$$u_t = B^\dagger \Big(M(q_t)\partial_t^2 p_t + C(q_t, \partial_t q_t)\partial_t q_t + G(q_t) - \hat{d}(q_t, \partial_t q_t) \Big). \tag{376}$$

7.1.1.2 Single drone – drag

In the presence of ambient wind and air resistance the state space equations of eq. (372) get an additional term

$$\begin{aligned} \partial_t p_t &= v_t & m\partial_t v_t &= mg + R_t u_t^f + d^f(x_t, \partial_t x_t, c_t) \\ \partial_t R_t &= R[\omega]_\times & J\partial_t \omega_t &= J\omega \times \omega + u_t^\tau + d^\tau(x_t, \partial_t x_t, c_t) \end{aligned} \tag{377}$$

where d^f and d^τ describe the forces and torques, respectively, of the wind with speed c_t and the air resistance. d^f and d^τ can be interpreted as a disturbance $d(q_t, \partial_t q_t)$. Hence, to get rid of its effect on the drone we need a good estimators $\hat{d}^f$ and $\hat{d}^\tau$ of d^f and d^τ respectively. The drag force is a nonlinear function of the speed of the drone relative to the wind and physical properties of the drone like the area of the drone exposed to the wind. This drag can be modelled using fluid dynamics,

$$\begin{aligned} \hat{d}^f(x_t, \partial_t x_t, c_t)_i &= c_i^f((v_t - c_t)_i)^2 \\ \hat{d}^\tau(x_t, \partial_t x_t, c_t)_i &= c_i^\tau((v_t - c_t)_i)^2 \end{aligned} \tag{378}$$

for $i \in \{0, 1, 2\}$. This approach will require fitting the six parameters c_i^f and c_i^τ. These values can be found by placing the drone in a wind tunnel. In practice the wind speed will be unknown, though. This will require an estimate of c_t. One way to solve this is using an adaptive control law. In this the parameters of the controller are changed based on the response of the drone to the control signal. This is done e.g. by comparing the expected behaviour of the drone based on the controller with the actual behaviour and estimating what c_t must have been. The controller can then be adapted to take into account the effect of what c_t was and update the control law to take into account possible future values of c_t. We will not go deeper into what a good adaptive control law would be for this problem, since this is not related to how deep learning is involved in control.

The performance of the adaptive control law does rely on whether $\hat{d}^f$ and $\hat{d}^\tau$ are good models for the physics behind the problem. The equations of eq. (378) can be argued to not be good models. The c_i^f and c_i^τ will only be a good fit for certain ranges, since they are dependent on v_t and c_t too. Furthermore, d_i^f and d_i^τ are not necessarily quadratic. To find good models $\hat{d}^f$ and $\hat{d}^\tau$, we can use neural networks.

Neural networks n^f and n^τ can made to find a good approximation of d^f and d^τ by finding parameters Θ^f and Θ^τ such that

$$\begin{aligned} n^f(x_t, \partial_t x_t, c_t; \Theta^f) &\approx d^f(x_t, \partial_t x_t, c_t) \\ n^\tau(x_t, \partial_t x_t, c_t; \Theta^\tau) &\approx d^\tau(x_t, \partial_t x_t, c_t) \end{aligned} \tag{379}$$

by minimizing

$$\min_{\Theta^f} \sum_{i=1}^{m} (n^f(x_i, y_i, c_i; \Theta^f) - d_i^f)^2 \tag{380}$$

and

$$\min_{\Theta^\tau} \sum_{i=1}^{m} (n^\tau(x_i, y_i, c_i; \Theta^\tau) - d_i^\tau)^2 \tag{381}$$

for m data points (x_i, y_i, c_i), d_i^τ and d_i^f representing $(x_t, \partial_t x_t, c_t)$, $d^f(x_t, \partial_t x_t, c_t)$ and $d^\tau(x_t, \partial_t x_t, c_t)$ collected at different times. Since neural networks are universal approximators, they will be able to estimate d^f and d^τ accurately.

To use the neural network in practice, knowledge of x_t, $\partial_t x_t$, and c_t is required. x_t and $\partial_t x_t$ are known, but c_t is not. This means that to use the neural networks n^f and n^τ in their current form requires measuring or estimating c_t. It is also possible to change the form of the neural network. Instead of finding neural networks of the form $n^f(x_t, \partial_t x_t, c_t; \Theta^f)$, it also possible to look for neural networks of the form $\tilde{n}^f(x_t, \partial_t x_t; \Theta^f)$ minimizing

$$\min_{\Theta^f} \min_{a \in \mathbb{R}} \sum_{i=1}^{m} (\tilde{n}^f(x_i, y_i; \Theta^f) \odot a - d_i^f)^2 \tag{382}$$

and

$$\min_{\Theta^\tau} \min_{a \in \mathbb{R}} \sum_{i=1}^{m} (\tilde{n}^\tau(x_i, y_i; \Theta^\tau) \odot a - d_i^\tau)^2, \tag{383}$$

where $\odot$ means pointwise multiplication. The inner minimisation problems of equations 382 and 383 are least squares for fixed Θ^f and Θ^τ respectively. Hence, once the neural network parameters Θ are determined based on collected data points, an adaptive control law can be determined that uses least squares to set a^f and a^τ such that

$$\begin{aligned} \hat{d}^f(x_t, \partial_t x_t, c_t) &= \tilde{n}^f(x_i, y_i; \Theta^\tau) \odot a^f \approx d^f(x_t, \partial_t x_t, c_t), \\ \hat{d}^\tau(x_t, \partial_t x_t, c_t) &= \tilde{n}^\tau(x_i, y_i; \Theta^\tau) \odot a^\tau \approx d^\tau(x_t, \partial_t x_t, c_t). \end{aligned} \tag{384}$$

For the proof and a possible adaptive control law see [O'Connell et al., 2021].

7.1.1.3 Single drone – ground effect

The drone has to fly over terrain and has to avoid obstacles. When it gets close to the ground or an obstacle, its flight characteristics change. The proximity changes how air flows pull or push the

drone in a certain direction. This requires a different disturbance model than that from section 7.1.1.2.

In section 7.1.1.2 the main challenge was the fact that c_t is unknown. In this case we know x_t and the surrounding map. However, there is no good estimate available for the effect of certain terrain on the drone. Only for some terrain is it possible to write an equation that describes the interaction between the drone and the terrain [Kan et al., 2019]. This means that we should not expect to be able to manually find a good model for the varying effects the shape, textures and proximity of the terrain and obstacles have on the drone. On the other hand, we can find a good approximation of the disturbance using a neural network, since x_t, $\partial_t x_t$ and the terrain are assumed to be known.

The terrain will be known to the drone in the form of a mesh map. This mesh map will include many points. Although neural networks can be used to approximate any function, it still takes time to evaluate that function. This means that trying to approximate a function that takes the entire mesh together with x_t and $\partial_t x_t$ as input might be to slow to evaluate. Instead of trying to approximate a function that takes the entire mesh, it may be more useful to only consider the mesh points within a certain distance from the drone as input.

7.1.1.4 Drone swarm – downwash

When two or more drones fly close to each other, they disturb each other. This effect is similar to the ground effect but in the opposite direction. If a drone flies close to the ground, it gains some lift. If the drone flies below another drone, it is pushed down [at Caltech, 2020]. This is highly dependent on the proximity, size, and motion of the drones. It is hard to find a good model for this disturbance [Modeling of aerodynamic disturbances for proximity flight of multi-rotors]. Since the drones are able to communicate their respective x_t and $\partial_t x_t$ to each other and can be made aware of their respective specs, it is again possible to find a good approximation of the disturbance using a neural network.

7.1.2 Approximate dynamic programming

In the previous section we looked at model adaptive control; control based on physics, reliant on suitable models for processes, with control laws that could adapt to instantaneous and exogenous information. In this section we will look at a model free method called approximate dynamic programming, in the machine learning community known by the name *reinforcement learning*. In this method an *actor* takes *actions* at each time step and *observes* the *environment* afterwards and gains a *reward* from its action. This reward does not have to be positive. In our example the drone represents the actor, the movement of the drone is an action and the reward it receives is, among others, the battery power it used to perform the action. In the method the actor wants to get the maximal reward at each time step. This means that for each observation o_t of the environment it will take the available action a_t that gives the highest reward r_t. The actor learns what actions to take based on a value function

$$Q : \mathfrak{O} \times \mathfrak{A} \to \mathfrak{R}, \tag{385}$$

where $\mathfrak{O}$ is the set of all observations of the environment, $\mathfrak{A}$ is the set of all possible actions, and $\mathfrak{R}$ is the set of rewards. If the environment is observed to be in state o_t, then the best action for the actor to take is

$$a_t^\dagger = \arg\max_{a \in \mathfrak{A}} Q(o_t, a). \tag{386}$$

We want to find a Q with suitable properties such that the actor takes desirable actions. This is done by constructing rewards smartly, and then iteratively updating the map Q based on new experiences in the form of observation-action-reward triples $(o_t, a_t, r_t) \in \mathfrak{O} \times \mathfrak{A} \times \mathfrak{R}$. The way that Q is updated differs between the solving methods. In all but the simplest of cases $\mathfrak{O} \times \mathfrak{A}$ is too large to iterate. Hence, solving methods typically revolve around sampling or pruning $\mathfrak{O} \times \mathfrak{A}$ in a way that the estimated Q is close to the real Q.

Remark. *We talk about observation of the state of the environment o_t and not about the state of the environment. This is because the actor typically is not able to observe the entire environment but just a fraction of it.*

A reason for not using approximate dynamic programming, outside of PID control, is that $\mathfrak{O} \times \mathfrak{A}$ is too large, and the methods available are not strong enough to give good results. By approximating Q with a neural network it becomes possible to use approximate dynamic programming in almost every control problem outside of system identification problems. To exemplify how this works in a control setting, we will again use the example of the drone swarm.

7.1.2.1 Observations and Actions

Let us start by defining the observation set $\mathfrak{O}$ and the actions set $\mathfrak{A}$. Recall that the drone has a state $x_t = (p_t, v_t, R_t, \omega_t)$ and its controller gives $u_t = (u_t^f, u_t^\tau)$. From this it follows that

$$\mathfrak{A} = \mathcal{U} = \{\text{possible controls } u_t\} \subseteq \mathbb{R}^3 \times \mathbb{R}^3. \tag{387}$$

For the observation set note that the drone is able to observe its own state x_t, but it also receives information about the states of the surrounding drones and terrain. Hence,

$$\mathfrak{O} = \left(\mathbb{R}^3 \times \mathbb{R}^3 \times SO(3) \times \mathbb{R}^3\right)^N \times \mathfrak{G} \tag{388}$$

where $\mathfrak{G}$ is the set of all terrain meshes.

The size of $\mathfrak{A}$ is limited, since the forces and torques that the drone can deliver are bounded. Futhermore, the 6 parameters are determined by how fast the 4 motors can spin. Hence, 4 parameters would also be sufficient. Since we need to take the argmin over all the actions, it is needed to discretize $\mathfrak{A}$. If the discretization is fine enough, then this will not noticeably degrade performance.

The size of $\mathfrak{O}$ is too big for practical purposes when N gets too large. One way of shrinking $\mathfrak{O}$ is by considering only the closest K drones with $K \ll N$ in an observation. Then $\mathfrak{O}$ becomes

$$\mathfrak{O} = \left(\mathbb{R}^3 \times \mathbb{R}^3 \times SO(3) \times \mathbb{R}^3\right)^{K+1} \times \mathfrak{G}. \tag{389}$$

7.1.2.2 Rewards

Recall that we want these drones to fly to their desired location without bumping into each other, without crashing and without taking an unnecessarily long route. We need to design rewards in such a way that this is achieved. One way of doing this is by setting the reward r_t to

$$r_t = c_{loc} r_t^{(loc)} + c_{col} r_t^{(col)} + c_{rel} r_t^{(rel)} + c_{terrain} r_t^{(terrain)} + c_{pow} r_t^{(pow)}, \tag{390}$$

where $r_t^{(loc)}$ is the reward for the drones proximity to the desired location, $r_t^{(rel)}$ is the reward for its relative location to the other drones, $r_t^{(terrain)}$ is the reward for its interaction with the terrain, $r_t^{(pow)}$ is the reward for using less battery power, and c_{loc}, c_{rel}, $c_{terrain}$ and c_{pow} are coefficients to balance the influences of the rewards on the total reward.

The drone is flying through Euclidian space, so a natural way to measure the distance between the drone and its target is the ℓ^2 norm. This means that the reward

$$r_t^{(loc)} = \left\| p_t - p_t^{(target)} \right\|_{\ell^2} \tag{391}$$

where $p_t^{(target)}$ is the target at time t, incentivizes to get close to the target. This reward is linear with the distance. It is possible to raise the norm to some higher power to reward the drone for flying closer.

Although the drone should fly to its target directly, it should not crash whilst doing so. Hence, all crashes should be penalised. A suitable reward for this is

$$r_t^{(col)} = \begin{cases} -1 & \text{crashed,} \\ 0 & \text{otherwise.} \end{cases} \tag{392}$$

Flying too close to other drones may not directly lead to crashes, but it should be disincentivized, because the downwash and changes in c_t may require the drone to make impossible manoeuvres. However, only the drones that are close are relevant. This means that the reward should vanish when the other drones are far enough away. This motivates a reward of the form

$$r_t^{(rel)} = -\sum_{j \neq i, j=1}^{N} \max\{1 - \frac{\left\| p_t^i - p_t^j \right\|_{\ell^2}}{d_{drop-off}^{(rel)}}, 0\}, \tag{393}$$

where p_t^k is the position of drone k at time t and $d_{drop-off}^{(rel)} > 0$ representing the distance after which we deem the influence of the other drone irrelevant.

Flying too close to the ground is not much different from flying too close to other drones. However, we might want to promote staying far away from the ground in general. Hence,

$$r_t^{(terrain)} = -\max_i \max\{1 - \frac{\|p_t - g_i\|_{\ell^2}}{d_{drop-off}^{(rel)}}, 0\}, \tag{394}$$

where g_i are the mesh points of the local terrain known by the drones.

We have assigned rewards for each task of the drone, but the drone is not yet incentivized to be gentle with its battery and motors. One way of achieving this is by penalizing high controls, i.e.

$$r_t^{(pow)} = -c_f \left\| u_t^f \right\|_{\ell^2} - c_\tau \| u_t^\tau \|_{\ell^2} \tag{395}$$

where c_f and c_τ determine the relative contribution of the forces and torques respectively.

Note that these are not the only ways to design rewards for the drone. For example when setting a value for $r_t^{(rel)}$ we set a radially symmetric reward. The drones affect each other differently horizontally and vertically, so it might be valuable to design a reward that is different in different directions. Also observe that it is possible to instruct the drone to hover near an object with these rewards. If the drones should hover near a ball hanging from a ceiling, then the target $p^{(target)}$ can be set to the center of the ball and the relative sizes of c_{loc}, c_{col} and $c_{terrain}$ will ensure that the reward is maximized for hovering near the ball.

7.1.2.3 Learning Q

In essence what we have done so far is constructing the state, actions and rewards of a Markov decision process, albeit using different symbols and different terminology. We have seen that the observation set $\mathfrak{O}$ is large, and this means that the value function Q cannot be computed using classical method used for Markov decision processes. However, Q can be approximated sufficiently well by using deep learning methods. We will discuss how to approximate Q using deep Q learning.

The core of deep Q learning is a simple value function update based on the Bellman equation. First, Q is a neural network initialised with arbitrary weights and biases. Then, given an observation of the current environment o_t, a chosen action a_t, a reward r_t, and an observation of the environment in the next step, the update is given by

$$Q^{new}(o_t, a_t) = (1-\alpha) Q^{old}(o_t, a_t) + \alpha \left(r_t + \gamma \max_a Q^{old}(o_{t+1}, a) \right) \tag{396}$$

where $\alpha \in (0,1)$ is a parameter determining the relative importance of the current value of Q^{old} and the just gained reward, and $\gamma \in [0,1)$ is a factor that determines the relative importance of potential future rewards compared to immediate rewards.

Equation (396) is implemented in a neural network by setting the loss of the network to

$$\left((r_t + \gamma \max_a Q(o_{t+1}, a)) - Q(o_t, a_t) \right)^2 \tag{397}$$

and the learning rate to α. The architecture of the network is fixed in the input and output dimensions. The input dimension is determined by the shape of o_t, and the output dimension is determined by the size of $\mathfrak{A}$. The remaining architecture can be chosen arbitrarily.

Data of the form (o_t, a_t, r_t, o_{t+1}) are needed to train the neural network. These can be found by flying a drone or by simulating a drone flying. The choice for which is used depends on the problem. In this case flying a drone gives physically accurate data, but there is a risk of crashing a precious

drone. Simulating is safer, but requires computing power and simulation software capable of giving physically accurate data. Although Q learning does not require models for how physical processes work, the simulation software does require them.

7.1.3 Model-based and model-free control with deep learning

In the previous two subsections we showed how deep learning influences control by means of an example using model adaptive control and approximate dynamic programming. Model adaptive control is a model-based method and approximate dynamic programming is a model-free method. We have shown that deep learning comes in a different form in both methods.

In model adaptive control the controller is designed with some parameters that can be changed based on real time information. The design of the controller and the associated parameters depend on models of the physical processes involved. If there is no model for a process, then a model has to be fitted to the process. If the fit gives a nice or simple function or the effect of the process is small, then it can be handled in the design process. However, classical control methods struggle with high dimensional and highly nonlinear processes. By approximating the effects of those processes with a neural network, model adaptive control methods can be extended to cover more problems.

In approximate dynamic programming the control problem is cast as a Markov decision process. Instead of modelling physical processes and fitting parameters for those processes, an observation set $\mathfrak{O}$, action set $\mathfrak{A}$ and rewards functions are defined for which the relative importance of the various rewards is determined by setting coefficients manually. The action set $\mathfrak{A}$ for control problems can be approximated by a small enough number of elements, but the observation sets $\mathfrak{O}$ are typically too large to compute the value function Q. This means that it was not often used as a method for solving control problems. This has become a feasible solving method by approximating the value function Q with a neural network.

In both model adaptive control and approximate dynamic programming we have seen that learning methods are used as good function approximators to either improve existing methods or make previously impractical methods possible.

7.2 Control in Deep Learning

In the previous section we have seen how control uses deep learning. To see how deep learning in turn is being influenced by control theory, we consider an optimisation problem. This optimisation problem is one of the optimisation problems in the class called *Neural ODE*s. These Neural ODE optimisation problems are generalisations of the underlying optimisation problem when trying to find the right parameters for residual neural networks.

The optimisation problem we will consider is

$$\begin{aligned} \min_{\mu \in L^1([0,T], rca(\Omega))} \frac{1}{2}\|f - z_T\|^2_{L^2(\rho,\mathbb{R}^d)} + \alpha \int_0^T \|\mu_t\|_{rca(\Omega)} dt, \\ \partial_t z_t(x) = K^\Omega_\sigma \mu_t(z_t(x)), \\ z_0(x) = x, \end{aligned} \tag{398}$$

with $\Omega \subseteq \mathbb{R}^{d \times d}$ and $T > 0$. Recall that K^Ω_ϕ was given by

$$K^\Omega_\sigma : rca(\Omega) \to L^2(\rho, \mathbb{R}^d),\ \mu \mapsto \left(x \mapsto \int_\Omega \phi(Ax + b) d\mu(A, b) \right), \tag{399}$$

for some pointwise applied, monotonically increasing function $\phi \in C^{0,1}(\mathbb{R})$. Note that although the optimisation problem is phrased using *min*, we are interested in knowing whether there exists a solution, how big the L^2 term is, and what the optimal μ is. In this work we will take some small steps in that direction.

In section 7.2.1 we discuss what these residual neural networks are, and how their generalisation leads to eq. (398). The procedure used shows that residual neural networks can be seen as space and time discretizations of control problems. Then, in section 7.2.2 we use control techniques in the form of the Hamiltonian equations to compute necessary conditions for optimality. Finally, we will take a look at what kind of functions can be approximated with vanishing L^2 term in section 7.2.3.

7.2.1 ResNets, and how they generalize to Neural ODEs

Neural ODEs find their origin in the infinite depth limit of the *residual neural networks*, ResNet for short. Until now we have mostly dealt with shallow neural networks, i.e. functions of the form

$$f_m(x) = \sum_{i=1}^{m_1} c_i \phi(A_i x + b_i) \tag{400}$$

with m, c_i, A_i, and b_i integers, scalars, matrices and vectors respectfully of appropriate dimensions. Deep neural networks are L shallow neural networks, $L \in \mathbb{N}$ and $L > 1$, stuck back to back such that the output of one is used as the input of the next, i.e. functions of the form

$$\begin{aligned} z^1(x) &= \sum_{i=1}^{m_1} c_i^1 \phi(A_i^1 x + b_i^1) \\ z^{\ell+1}(x) &= \sum_{i=1}^{m_\ell} c_i^\ell \phi(A_i^\ell z^\ell(x) + b_i^\ell) \quad \ell \in \{1, ..., L-1\} \end{aligned} \tag{401}$$

with m_ℓ, c_i^ℓ, A_i^ℓ, and b_i^ℓ integers, scalars, matrices and vectors respectfully of appropriate dimensions. Many more functions can be approximated by doing these concatenations of shallow neural networks. The parameters c_i^ℓ, A_i^ℓ, and b_i^ℓ in the deep neural networks are updated using a form of gradient descent. That involves taking the gradient of z_L with respect to some parameter. The chain rule tells us that many terms need to be multiplied for some of these gradients. Multiplying many numbers can lead to vanishingly small numbers or humongously large numbers. This means that gradient

descent may cause sometimes parameters to stay the same or to blow up. These are undesirable things to happen. By using ResNets instead of the deep neural networks of eq. (401), these vanishing gradients can be avoided. ResNets do this by adding an additional term, i.e. they are functions of the form

$$\begin{aligned} z^1(x) &= x + \sum_{i=1}^{m_1} c_i^1 \phi(A_i^1 x + b_i^1) \\ z^{\ell+1}(x) &= z^\ell(x) + \sum_{i=1}^{m_\ell} c_i^\ell \phi(A_i^\ell z^\ell(x) + b_i^\ell) \quad \ell \in \{1, ..., L-1\} \end{aligned} \tag{402}$$

with m_ℓ, c^ℓ, A_i^ℓ, and b_i^ℓ integers, scalars, matrices and vectors respectfully of appropriate dimensions. Adding $z^\ell(x)$ to $z^{\ell+1}(x)$ causes the gradients to always have a term that is at most one multiplication of factors. This term is unlikely to vanish compared to the terms that have many multiplications of factors. The ResNet is thus a more stable version of the deep neural networks of eq. (401).

In the deep neural networks of eq. (401) and the ResNet the superscript ℓ represents the layers. By treating the layers as a time discretization eq. (402) can be seen as the Euler discretization of the ODE

$$\begin{aligned} z_0(x) &= x \\ \partial_t z_t(x) &= \sum_{i=1}^{m_t} (c_i)_t \phi((A_i)_t z_t(x) + (b_i)_t) \end{aligned} \tag{403}$$

with m_t, $(c_i)_t$, $(A_i)_t$, and $(b_i)_t$ now integer, scalar, matrix and vector valued functions of time respectfully. ODEs of the form given in eq. (403) are called *Neural ODE*s.

Recall that to get the infinite width limit of the shallow neural network we replaced the sum of eq. (400) with an integral and moved c_i into the measure $\mu \in rca(\Omega)$ over which is integrated. Neural ODEs still have the sum structure, but parameters now vary in time. This means the measure should also be time varying. Hence, we can write the infinite width limit of eq. (403) as

$$\begin{aligned} z_0(x) &= x \\ \partial_t z_t(x) &= \int_\Omega \phi(A z_t(x) + b) d\mu_t(A, b) = K\mu_t(z_t(x)) \end{aligned} \tag{404}$$

with $\mu \in L^1([0, T], rca(\Omega))$. Equation (404) is the state space equation used in eq. (398).

We still need to show why we minimise over what we minimise in eq. (398). Recall that the Barron norm was given by

$$\|f\|_{\mathcal{B}_\phi^\Omega} = \inf_{\mu \in M_{\phi,f}^\Omega} \int_\Omega W_\phi(A, b) d|\mu|(A, b), \tag{405}$$

where W_ϕ represents the weights given to A and b and $M_{\phi,f}^\Omega$ is the set of all measures μ such that $K_\phi^\Omega \mu = f$. The L^2 relaxation of eq. (405) is

$$\begin{aligned} \inf_{\mu \in rca(\Omega)} \frac{1}{2} \|f - z\|_{L^2(\rho, \mathbb{R}^d)}^2 + \alpha \int_\Omega W_\phi(A, b) d|\mu|(A, b) \\ z(x) = K_\phi^\Omega \mu(x). \end{aligned} \tag{406}$$

When we replace the $z = K_\sigma^\Omega \mu$ of eq. (406) by eq. (404), we need to change the weight term

$$\int_\Omega W_\sigma(A, b) d|\mu|(A, b) \tag{407}$$

such that time is properly taking into account. One way to do this is simply by integrating, i.e. by changing eq. (407) to

$$\int_0^T \int_\Omega W_\sigma(A,b) d|\mu_t|(A,b) dt. \tag{408}$$

Adapting eq. (406) with eq. (404) and eq. (408) gives

$$\begin{aligned} \min_{\mu_t \in C([0,T], rca(\Omega))} & \frac{1}{2}\|f - z_T\|^2_{L^2(\rho,\mathbb{R}^d)} + \alpha \int_0^T \int_\Omega W_\sigma(A,b) d|\mu_t|(A,b) dt, \\ & \partial_t z_t(x) = K^\Omega_\phi \mu_t(z_t(x)), \\ & z_0(x) = x. \end{aligned} \tag{409}$$

If we take the Bach version of W_σ, i.e. remove W_σ from eq. (409) and replace it with $W : (A,b) \mapsto 1$, then we get eq. (398).

7.2.2 Hamiltonian Equations

The procedure of section 7.2.1 shows that ResNets, and by extension the deep neural networks of eq. (401), are in fact space and time discretizations of an optimal control problem. This brings us the question: If we apply ideas from control theory to eq. (398), can we learn something new about ResNets? In this section we investigate this question by looking at the Hamiltonian equations. To simplify the notation, we omit the super- and subscript from K^Ω_ϕ, i.e. we write K instead of K^Ω_ϕ.

The Hamiltonian for eq. (398) is given by

$$H(\mu_t, z_t, p_t) = \langle p_t | K\mu_t \circ z_t \rangle_{L^2(\rho)} + \alpha \|\mu_t\|_{rca(\Omega)}. \tag{410}$$

The Hamiltonian equations,

$$\begin{aligned} \partial_t z_t &= \partial_{p_t} H(\mu_t, z_t, p_t), \quad z_0(x) = x, \\ \partial_t p_t &= -\partial_{z_t} H(\mu_t, z_t, p_t), \quad p_T = \partial_{z_T} \frac{1}{2}\|f - z_T\|^2_{L^2(\rho)}, \\ 0 &= \partial_{\mu_t} H(\mu_t, z_t, p_t), \end{aligned} \tag{411}$$

describe the first order optimality conditions for eq. (398). For μ_t and z_t to be optimal solutions there must be a p_t such that p_t, μ_t and z_t together solve eq. (411). We will now compute the derivatives of eq. (411) to see what these optimality conditions imply.

Proposition 7.1. *Let*

$$K' : rca(\Omega) \to L^2(\rho, \mathbb{R}^d),\ \mu \to \left(x \mapsto \int_\Omega \partial\phi(Ax+b) A d\mu(A,b) \right), \tag{412}$$

$$K^* : L^2(\rho,\mathbb{R}^d) \times L^2(\rho,\mathbb{R}^d) \to C(\Omega),\ (A,b) \mapsto \left((g,h) \mapsto \int_{\mathcal{X}} \langle g(x) | \phi(Ah(x)+b) \rangle_{\ell^2} \, d\rho(x) \right), \tag{413}$$

then the Hamiltonian equations for eq. (398) *can be represented as*

$$\begin{aligned} \partial_t z_t &= K\mu_t \circ z_t, \quad z_0(x) = x, \\ \partial_t p_t &= -\left\langle p_t \middle| K'\mu_t \circ z_t \right\rangle_{L^2(\rho,\mathbb{R}^d)}, \quad p_T = z_T - f, \\ \operatorname{sgn}\{\mu_t\} &= \frac{-1}{\alpha} K^*(p_t, z_t) \quad \mu \text{ a.e.}. \end{aligned} \tag{414}$$

Proof. To show that eq. (414) is equal to eq. (411) we have to compute 4 derivatives.

The first is immediate,

$$\partial_{p_t} H(\mu_t, z_t, p_t) = K\mu_t \circ z_t. \tag{415}$$

The second follows from an application of the chain rule and the fact that p_t as well as μ_t do not directly depend on z_t,

$$\begin{aligned}
\partial_{z_t} H(\mu_t, z_t, p_t) &= \partial_{z_t}\Big(\langle p_t | K\mu_t \circ z_t \rangle_{L^2(\rho)} + \alpha \|\mu_t\|_{rca(\Omega)} \Big) \\
&= \langle p_t | \partial_{z_t} (K\mu_t \circ z_t) \rangle_{L^2(\rho)} \\
&= \left\langle p_t \middle| \partial_{z_t} \int_\Omega \phi(Az_t(x) + b) d\mu_t(A, b) \right\rangle_{L^2(\rho)} \\
&= \left\langle p_t \middle| \int_\Omega \partial_{z_t} \phi(Az_t(x) + b) d\mu_t(A, b) \right\rangle_{L^2(\rho)} \\
&= \left\langle p_t \middle| \int_\Omega \partial\phi(Az_t(x) + b) A d\mu_t(A, b) \right\rangle_{L^2(\rho)} \\
&= \left\langle p_t \middle| K'\mu_t \circ z_t \right\rangle_{L^2(\rho)}.
\end{aligned} \tag{416}$$

The third follows from the Fréchet derivative of the L^2 norm,

$$\partial_{z_T} \frac{1}{2} \|f - z_T\|^2_{L^2(\rho)} = z_T - f. \tag{417}$$

For the fourth and last we need three intermediate results first. These are that by Fubini

$$\langle K\mu \circ h | g \rangle_{L^2(\rho, \mathbb{R}^d)} = \langle K^*(g, h) | \mu \rangle_{rca(\Omega)} \tag{418}$$

for all $\mu \in rca(\Omega)$ and $f, g \in L^2(\rho, \mathbb{R}^d)$, that there must be a function $\operatorname{sgn}\{\mu_t\} \in L^1(\mu)$ with $|\operatorname{sgn}\{\mu_t\}| = 1$ μ a.e. such that

$$d|\mu| = \operatorname{sgn}\{\mu_t\} d\mu \tag{419}$$

since $|\mu| \ll \mu$, and that

$$\partial_\mu \langle g | \mu \rangle_{rca(\Omega} = g \tag{420}$$

for all $\mu \in rca(\Omega)$ and $g \in L^1(\mu)$ since

$$\lim_{\|\nu\|_{rca(\Omega)} \to 0} \frac{\left| \langle g | \mu + \nu \rangle_{rca(\Omega} - \langle g | \mu \rangle_{rca(\Omega)} - \left\langle \partial_\mu \langle g | \mu \rangle_{rca(\Omega} \middle| \nu \right\rangle_{rca(\Omega)} \right|}{\|\nu\|_{rca(\Omega)}} = 0. \tag{421}$$

With these we can write

$$\begin{aligned}
\partial_{\mu_t} H(\mu_t, z_t, p_t) &= \partial_{\mu_t}\Big(\langle p_t | K\mu_t \circ z_t \rangle_{L^2(\rho, \mathbb{R}^d)} + \alpha \|\mu_t\|_{rca(\Omega)} \Big) \\
&= \partial_{\mu_t}\Big(\langle K^*(p_t, z_t) | \mu_t \rangle_{rca(\Omega)} + \alpha \|\mu_t\|_{rca(\Omega)} \Big) && eq.\ (418) \\
&= K^*(p_t, z_t) + \alpha \partial_{\mu_t} \|\mu_t\|_{rca(\Omega)} && eq.\ (419) \\
&= K^*(p_t, z_t) + \alpha \operatorname{sgn}\{\mu_t\}. && eq.\ (420)
\end{aligned} \tag{422}$$

Substituting equations 415 till 422 into eq. (411) gives eq. (414). *Q.E.D.*

The third equation of eq. (414) is an interesting result. Since

$$|\mathrm{sgn}\{\mu_t\}| = 1 \quad \mu_t \text{ a.e.}, \tag{423}$$

it must hold that

$$\operatorname{supp}\mu_t \subseteq \left\{ (A,b) \in \Omega \;\middle|\; |K^*(p_t,z_t)(A,b)| = \alpha \right\}. \tag{424}$$

Additionally, the function K^*p_t is a continuous function. This means it is limited in how fast it changes sign, and by extension how fast μ_t can. Combined with eq. (424) this implies that the sets $\operatorname{supp}(\mu_t)_+$ and $\operatorname{supp}(\mu_t)_-$, the sets where μ_t takes positive and negative values respectively, must be separated by a non-zero distance. It, however, does not tell us what that non-zero distance is. By construction of K and K^* we can provide an estimate. For that recall that a lipschitz function on g on Ω satisfies

$$\left| g(A,b) - g(\tilde{A},\tilde{b}) \right| \leqslant Lip(g) \left\| (A,b) - (\tilde{A},\tilde{b}) \right\|. \tag{425}$$

If we know how far $g(A,b)$ must be from $g(\tilde{A},\tilde{b})$ and we know $Lip(g)$, then we have a lower bound on $\left\|(A,b) - (\tilde{A},\tilde{b})\right\|$. We will now use this idea to compute an estimate for the minimal distance between the sets $\operatorname{supp}(\mu_t)_+$ and $\operatorname{supp}(\mu_t)_-$.

Proposition 7.2. *For μ_t, z_t and p_t satisfying eq.* (414) *it must hold that*

$$\inf\left\{ \left\| (A,b) - (\tilde{A},\tilde{b}) \right\| \;\middle|\; (A,b) \in \operatorname{supp}(\mu_t)_+, (\tilde{A},\tilde{b}) \in \operatorname{supp}(\mu_t)_- \right\} \geqslant \frac{2\alpha}{Lip(K^*(p_t,z_t))}. \tag{426}$$

Proof. From eq. (424) it follows that

$$K^*(p_t,z_t)(A,b) = \pm\alpha \tag{427}$$

for $(A,b) \in \operatorname{supp}(\mu_t)_\pm$. Hence,

$$\begin{aligned} 2\alpha &= |\alpha - (-\alpha)| \\ &= \left| K^*(p_t,z_t)(A,b) - K^*(p_t,z_t)(\tilde{A},\tilde{b}) \right| \quad \textit{eq. (427)} \\ &\leqslant Lip(K^*(p_t,z_t)) \left\| (A,b) - (\tilde{A},\tilde{b}) \right\| \end{aligned} \tag{428}$$

for all $(A,b) \in \operatorname{supp}(\mu_t)_+, (\tilde{A},\tilde{b}) \in \operatorname{supp}(\mu_t)_-$. If $Lip(K^*(p_t,z_t)$ is finite, then eq. (428) can be rewritten to

$$\left\| (A,b) - (\tilde{A},\tilde{b}) \right\| \geqslant \frac{2\alpha}{Lip(K^*(p_t,z_t))}. \tag{429}$$

Taking the infimum on the left hand side of eq. (429) over all $(A,b) \in \operatorname{supp}(\mu_t)_+, (\tilde{A},\tilde{b}) \in \operatorname{supp}(\mu_t)_-$ gives eq. (426).

What remains to show is that $Lip(K^*(p_t,z_t))$ is indeed finite. For that observe that

$$\begin{aligned} \left| K^*(p_t,z_t)(A,b) - K^*(p_t,z_t)(\tilde{A},\tilde{b}) \right| &= \left| \int_{\mathcal{X}} \langle p_t(x) | \phi(Az_t(x)+b) \rangle_{\ell^2} \, d\rho(x) - \int_{\mathcal{X}} \left\langle p_t(x) \middle| \phi(\tilde{A}z_t(x)+\tilde{b}) \right\rangle_{\ell^2} d\rho(x) \right| \\ &= \left| \int_{\mathcal{X}} \langle p_t(x) | \phi(Az_t(x)+b) \rangle_{\ell^2} - \left\langle p_t(x) \middle| \phi(\tilde{A}z_t(x)+\tilde{b}) \right\rangle_{\ell^2} d\rho(x) \right| \end{aligned}$$

$$
\begin{aligned}
&= \left| \int_{\mathcal{X}} \left\langle p_t(x) \middle| \phi(Az_t(x)+b) - \phi(\tilde{A}z_t(x)+\tilde{b}) \right\rangle_{\ell^2} d\rho(x) \right| \\
&\leqslant \int_{\mathcal{X}} \left| \left\langle p_t(x) \middle| \phi(Az_t(x)+b) - \phi(\tilde{A}z_t(x)+\tilde{b}) \right\rangle_{\ell^2} \right| d\rho(x) \\
&\leqslant \int_{\mathcal{X}} \|p_t(x)\|_{\ell^2} \left\| \phi(Az_t(x)+b) - \phi(\tilde{A}z_t(x)+\tilde{b}) \right\|_{\ell^2} d\rho(x) \\
&\leqslant Lip(\phi) \int_{\mathcal{X}} \|p_t(x)\|_{\ell^2} \left\| (Az_t(x)+b) - (\tilde{A}z_t(x)+\tilde{b}) \right\|_{\ell^2} d\rho(x) \\
&= Lip(\phi) \int_{\mathcal{X}} \|p_t(x)\|_{\ell^2} \left\| (A-\tilde{A})z_t(x) + (b-\tilde{b}) \right\|_{\ell^2} d\rho(x) \\
&\leqslant Lip(\phi) \int_{\mathcal{X}} \|p_t(x)\|_{\ell^2} \|1+z_t\|_{\ell^2} d\rho(x) \left\| (A-\tilde{A}) + (b-\tilde{b}) \right\|_{\ell^2} \\
&\leqslant Lip(\phi) \left(\|p_t\|^2_{L^2(\rho,\mathbb{R}^d)} + \|1+z_t\|^2_{L^2(\rho,\mathbb{R}^d)} \right) \left\| (A-\tilde{A}) + (b-\tilde{b}) \right\|_{\ell^2}
\end{aligned}
$$

for all $(A,b),(\tilde{A},\tilde{b}) \in \Omega$. Both p_t and z_t are $L^2(\rho,\mathbb{R}^d)$ functions and $\phi \in C^{0,1}(\Omega)$ thus Lipschitz, so $Lip(K^*(p_t,z_t))$ must be finite. *Q.E.D.*

Proposition 7.2 shows that the distance between the supports of $(\mu_t)_-$ and $(\mu_t)_+$ scales explicitly linearly with α. Unfortunately, $Lip(K^*p_t)$ implicitly depends on α, so the usefulness of this bound is debatable.

7.2.3 Functions that can be approximated

In most of this work we have looked at the Barron space, and what kind of properties the functions in that space have. One of the ways that we have done so is by looking at embeddings into other spaces. In this section we will do the same for the Neural ODE of eq. (404).

Consider the set of functions

$$
\mathfrak{ODE}^{\Omega}_{\phi} = \left\{ f \;\middle|\; \exists \mu \in L^1([0,T], rca(\Omega)) : z_T = f, z_0(x) = x, \partial_t z_t = K^{\Omega}_{\phi}\mu_t \circ z_t \right\}. \tag{430}
$$

This set represents the set of all functions that can be made using the Neural ODE. A natural question to start with is the question: Is this a vector space? The answer is unknown to the author. Due to the highly nonlinear and recursive nature of the ODE, it is generally not possible to get $f+g \in \mathfrak{ODE}$ by summing the two measures μ and ν associated to some $f,g \in \mathfrak{ODE}$. It is therefore not clear how to show that $c_1 f + c_2 g \in \mathfrak{ODE}$ when $f,g \in \mathfrak{ODE}$ and $c_1,c_2 \in \mathbb{R}$.

Although we have not been able to show whether $\mathfrak{ODE}$ can be made into a vector space, we can show two inclusions. The first shows that the functions in $\mathfrak{ODE}^{\Omega}_{\phi}$ must be Lipschitz, just like the Barron functions.

Proposition 7.3. *If $\phi \in C^{0,1}(\mathbb{R})$ is an activation function, then*

$$
\mathfrak{ODE}^{\Omega}_{\phi} \subseteq C^{0,1}(\mathcal{X}). \tag{431}
$$

Proof. Consider a function $f \in \mathfrak{ODE}$. There must be $z \in C^1([0,T], L^2(\rho, \mathbb{R}^d))$ such that $z_T = f$. This z satisfies

$$\begin{aligned}
\|z_t(x) - z_t(y)\|_{\ell^2} &= \left\| \int_0^t K_\phi^\Omega \mu_s(z_s(x))ds - \int_0^t K_\phi^\Omega \mu_s(z_s(x))ds \right\|_{\ell^2} \\
&\leqslant \int_0^t \left\| K_\phi^\Omega \mu_s(z_s(x))ds - \int_0^t K_\phi^\Omega \mu_s(z_s(x)) \right\|_{\ell^2} ds \\
&\leqslant \int_0^t \left\| K_\phi^\Omega \mu_s(z_s(x)) - K_\phi^\Omega \mu_s(z_s(x)) \right\|_{\ell^2} ds \\
&\leqslant \int_0^t Lip(K_\phi^\Omega \mu_s) \|z_s(x) - z_s(x)\|_{\ell^2} ds \qquad \textit{proposition } 6.1
\end{aligned} \tag{432}$$

for all $t \in [0,T]$. An application of Grönwall's inequality gives

$$\|z_t(x) - z_t(y)\|_{\ell^2} \leqslant \|z_0(x) - z_0(y)\|_{\ell^2} e^{\int_0^t Lip(K\mu_s)ds} = \|x - y\|_{\ell^2} e^{\int_0^t Lip(K\mu_s)ds}. \tag{433}$$

Equation (433) implies that z_t is Lipschitz for all $t \in [0,T]$. In particular, z_T is Lipschitz and, by extention, f. *Q.E.D.*

The second inclusion is the other way around. We show that certain affine functions can be approximated using a Neural ODE.

Proposition 7.4. *If $Q, P \in L^1([0,T])$, then $f \in \mathfrak{ODE}$ where*

$$f(x) = e^{-\int_0^T P_s ds} \left(x + \int_0^T e^{\int_0^s P_\tau d\tau} Q_s ds \right). \tag{434}$$

Proof. From proposition 3.4 it follows that there exists measures $\nu, \gamma \in rca(\Omega)$ such that

$$\begin{aligned}
K\nu(x) &= 1, \\
K\gamma(x) &= x.
\end{aligned}$$

Combined with the assumption on Q and P, we get that $\mu_t = Q_t\nu - P_t\gamma$ satisfies

$$\int_0^T \|\mu_t\|_{rca(\Omega)} dt \leqslant \|\nu\|_{rca(\Omega)} \int_0^T |Q_t| dt + \|\gamma\|_{rca(\Omega)} \int_0^T |P_t| dt. \tag{435}$$

Hence, $\mu \in L^1([0,T], rca(\Omega))$. Inserting μ into eq. (404) gives

$$\partial_t z_t(x) = -P_t z_t(x) + Q_t. \tag{436}$$

Solving eq. (436) for $z_t(x)$ gives

$$z_t(x) = e^{-\int_0^t P_s ds} \left(x + \int_0^t e^{\int_0^s P_\tau d\tau} Q_s ds \right). \tag{437}$$

From the combination of eq. (434) and eq. (437), it follows that $z_T = f$. Therefore, $f \in \mathfrak{ODE}$. *Q.E.D.*

Corollary 7.0.1. *$x \mapsto ax + b \in \mathfrak{ODE}$ for all $a > 0$ and $b \in \mathbb{R}$.*

What other functions there are in $\mathfrak{ODE}$ remains unknown, and in particular the question whether $\mathcal{B}_\phi^\Omega \subseteq \mathfrak{ODE}_\phi^\Omega$ remains elusive.

8 Future work and Open Questions

In this document we have furthered the understanding of Bach and Barron spaces, in particular for the higher order ReLU. However, questions remain. In this section some questions are discussed and, if applicable, conjectures are formulated. The questions are numbered and indented for clarity. After each question there is a small discussion of that question.

1. Let ϕ and ψ be two activation functions such that ϕ is the derivative of ψ. Does that mean that the Barron space with activation function ϕ embeds in the one with activation function ψ?

In section 3.4 we have discussed Barron spaces for various activation functions, mainly whether they embed into ReLU. The activation functions that we have considered are not an exhaustive list. Suppose, for example, that $\phi = \partial\psi$ for some activation functions ψ and ϕ. Recall that this means that

$$\phi(x) = \lim_{h\to 0} \frac{\psi(x+h) - \psi(x)}{h}. \tag{438}$$

This hints at the possibility of approximating the effect of a neuron with activation function ϕ up to arbitrary accuracy by using two neurons with activation function ψ. In section 3.1 it is shown that Barron space is complete, indicating that the sequences of pairs of neurons that approximate ϕ stay inside the Barron space. On the other hand, it must hold that

$$K^\Omega_\phi \mu(x) = \lim_{h\to 0} \frac{1}{h}\bigg(K^\Omega_\psi \mu(x+h) - K^\Omega_\psi \mu(x) \bigg). \tag{439}$$

If the factor within brackets goes to zero superlinearly in h, then it might be that the embedding of the Barron space with ϕ into the Barron space with ψ holds. This clearly puts a limit on which ψ this would work for. It is unknown to the author for which activation functions this process would yield an embedding.

2. How much larger is a direct sum of Barron spaces with activation functions ϕ_i compared to the Barron spaces with activation functions ϕ_i?

So far we have only considered Barron spaces with a single activation function. Each of these Barron spaces represents a set of functions. If we have two activation functions ϕ_i, then it is possible that both associated Barron spaces can approximate the same function f. Activation functions work on different scales. Hence, it might be possible that both the associated Barron norms are quite large. At the same time it is possible that a linear combination of functions $f_i \in B^\Omega_{\phi_i}$ can be used to represent f, i.e.

$$f(x) = c_1 f_1(x) + c_2 f_2(x) \tag{440}$$

for $c_i \in \mathbb{R}$, with much lower Barron norm, i.e.

$$c_1 \|f_1\|_{\mathcal{B}_{\phi_1}} + c_2 \|f_2\|_{\mathcal{B}_{\phi_2}} \ll \|f\|_{\mathcal{B}_{\phi_i}}. \tag{441}$$

Furthermore, we can always choose c_1, c_2, f_1, f_2 such that

$$c_1\|f_1\|_{\mathcal{B}^\Omega_{\phi_1}} + c_2\|f_2\|_{\mathcal{B}^\Omega_{\phi_2}} \leqslant \|f\|_{\mathcal{B}^\Omega_{\phi_i}} \tag{442}$$

for all $f \in \mathcal{B}_{\phi_i}$. This combined suggests that this space of linear combination of function $f_i \in \mathcal{B}_{\phi_i}$ contains more functions than each $\mathcal{B}_{\phi_i}$ and the total norm is lower. This is captured in the following proposition.

Proposition 8.1. *Let ϕ_i be activation functions, then*

$$\mathcal{B}^\Omega_{\phi_i} \hookrightarrow \bigoplus_i \mathcal{B}^\Omega_{\phi_i} \tag{443}$$

as well as

$$\bigcup_i \mathcal{B}^\Omega_{\phi_i} \subseteq \bigoplus_i \mathcal{B}^\Omega_{\phi_i}. \tag{444}$$

Proof. Follows directly from the definition of the direct sum $\bigoplus_i$. *Q.E.D.*

Although it might be interesting to know what kind of functions are in $\bigoplus_i \mathcal{B}_{\phi_i} \backslash \bigcup_i \mathcal{B}_{\phi_i}$, i.e. the set of functions that can be described by several activation functions but not with only one activation function, it would be more interesting to know how large the various errors of this space are. The author's guess is that the projection, approximation, and estimation errors will be similar or lower, but that the worst case training error is higher. The projection error is thought to be smaller, because the direct sum space contains more functions. The approximation and estimation errors are thought to be similar, because a similar construction as for the original Barron spaces can be used to bound them. The training error is thought to be larger, because the activation functions work on different scales possibly leading to more vanishing or exploding gradients.

3. Is it possible to formulate error bounds when the loss functional is changed from the $L^2(\mathcal{X}, \rho)$ norm to the Sobolev norm $H^s(\mathcal{X}, \rho)$?

In this work we have considered the L^2 loss. This does not include a term for the derivatives. In physical processes derivatives can be highly important. This means that we ideally also have a bound on how well we can approximate the derivatives. For the approximation error we can find inspiration in the works of Siegel and Xu. In [Siegel and Xu, 2021] the authors discuss the Sobolev loss $\mathcal{H}^s$, which does include derivatives albeit weak derivatives. In theorem 2 of their work they prove that, given sufficiently smooth and fast enough decaying activation functions, functions $f \in \mathscr{F}_I^{s+1,1}$ can locally be efficiently approximated in H^s by finite width shallow neural networks. In the proof of this theorem they explicitly construct a probability measure λ such that $f \in \mathscr{F}_I^{s+1,1}$ can be written as an expectation over λ. This strongly suggests that it is possible to construct, for each $f \in \mathscr{F}_I^{s+1,1}$, a measure μ with

$$\|\mu\|_{rca(\mathbb{R}^{d+1})} \leqslant C\|f\|_{\mathscr{F}_I^{s+1,1}} \tag{445}$$

for some $C > 0$ such that $f = K^\Omega_\phi \mu$. We make this more rigorous in the following conjecture.

Conjecture 8.1 (Adaptation of [Siegel and Xu, 2021; theorem 2 and corollary 1])**.** *Consider an activation function $\phi \in W^{s,\infty}(\mathbb{R})$ that is non-zero and satisfies the decay condition*

$$\left|\partial^k \phi(t)\right| \leqslant C_p(1+|t|)^{-p} \tag{446}$$

for $0 \leqslant k \leqslant s$ and some $p > 1$ and $C_p > 0$, then

$$\mathscr{F}_I^{s+1,1} \hookrightarrow V_\phi^{\mathbb{R}^{d+1}} \tag{447}$$

and

$$\|f - f_m\|_{H^s(\mathcal{X})} \leqslant C\frac{\|f\|_{V_\phi^{\mathbb{R}^{d+1}}}}{\sqrt{m}} \tag{448}$$

for $f \in \mathscr{F}_I^{s+1,1}$. If ϕ does not satisfy the decay condition but a linear combination ψ of $m_0 \in \mathbb{N}$ elements does, then instead

$$\mathscr{F}_I^{s+1,1} \hookrightarrow V_\psi^{\mathbb{R}^{d+1}} \hookrightarrow V_\phi^{\mathbb{R}^{d+1}} \tag{449}$$

and

$$\|f - f_m\|_{H^s(\mathcal{X})} \leqslant C\sqrt{m_0}\frac{\|f\|_{V_\phi}}{\sqrt{m}} \tag{450}$$

for $f \in \mathscr{F}_I^{s+1,1}$.

On page 18 of [Siegel and Xu, 2021] a table is given that shows that the higher-order ReLU satisfies the conditions of conjecture 8.1. This means that conjecture 8.1 implies in particular that $\mathscr{F}_I^{s,1} \hookrightarrow \mathcal{B}_{\sigma_s}^\Omega$, and thus fig. 10 too. Furthermore, conjecture 8.1 shows a bound for the approximation error. To provide a bound akin to that of theorem 4.4, it remains to be shown what the Rademacher complexity is and what the bounds for a direct approximation theorem are.

$$\begin{array}{ccccc} \mathscr{F}_I^{s+1,1} & \longrightarrow & \mathscr{F}_I^{t+1,1} & \longrightarrow & \mathscr{F}_I^{2,1} \\ & \searrow & & \searrow & \\ & \mathcal{B}_{\sigma_s}^\Omega & \longrightarrow & \mathcal{B}_{\sigma_t}^\Omega & \longrightarrow \mathcal{B}_{\sigma_1}^\Omega \end{array}$$

Figure 10: The relationship between $\mathcal{B}_{\sigma_k}^\Omega$ and $\mathscr{F}_I^{k+1,1}$ for $k \in \{s,t,1\}$ with $1 \leqslant t \leqslant s$. Valid if conjecture 8.1 holds. Arrows represent embeddings.

4. The authors of [Caragea et al., 2020] consider a space $\mathcal{B}_{\mathcal{F},s}(U)$. How does this space relate to $\mathscr{F}_I^{s,1}$ and the Barron spaces $\mathcal{B}_\phi^\Omega$?

In this work we have discussed various properties of the space $\mathscr{F}_I^{s,1}$. This function space places different restrictions on the functions in it. One of these restrictions is that the function must be defined over all or $\mathbb{R}^d$. However, often we are only interested in local properties. For example in the Fourier expansion of theorem 4.2 we are focus on the remainder of the Taylor expansion. This is a local property, but we use $\mathscr{F}_I^{s,1}$. One way to describe a local version of functions from $\mathscr{F}_I^{s,1}$ is to consider

$$\begin{aligned} \mathscr{F}_I^{s,1}(U) &= \left\{ f : U \to \mathbb{R} \;\middle|\; \|f\|_{\mathscr{F}_I^{s,1}(U)} < \infty \right\}, \\ \|f\|_{\mathscr{F}_I^{s,1}(U)} &= \inf\left\{ \|g\|_{\mathscr{F}_I^{s,1}} \;\middle|\; g \in \mathscr{F}_I^{s,1}, \forall x \in U : g(x) = f(x) \right\} \end{aligned} \tag{451}$$

for some $U \subset \mathbb{R}^d$. This can be interpreted as: if $f \in \mathscr{F}_I^{s,1}(U)$, then f must be able to be expanded to a function $g \in \mathscr{F}_I^{s,1}$. The norm of f is then determined by the smallest extension g. Clearly, $\mathscr{F}_I^{s,1} \subseteq \mathscr{F}_I^{s,1}(U)$. However, $\mathscr{F}_I^{s,1}(U) \nsubseteq \mathscr{F}_I^{s,1}$, since it is possible to have functions $f : \mathbb{R}^d \to \mathbb{R}$ that are not in $\mathscr{F}_I^{s,1}$ but have an extension such that they are in $\mathscr{F}_I^{s,1}(U)$. It is unclear whether in the propositions and theorems of this work $\mathscr{F}_I^{s,1}$ can be replaced by $\mathscr{F}_I^{s,1}(\mathcal{X})$ without any issue. In particular, theorem 4.2 would be more powerful. We conjecture that this is possible.

Conjecture 8.2. *If $f \in \mathscr{F}_I^{s+1,1}(\mathcal{X})$, then there exists an $R > 0$ and a measure $\mu \in rca(S^d \times [0, R])$ such that for all $x \in \mathcal{X}$*

$$f(x) = \sum_{|\beta| \leqslant s} \frac{1}{\beta!} \partial^\beta f(0) x^\beta + \tilde{f}(x) \tag{452}$$

with

$$\tilde{f}(x) = K_{\sigma_s}^{S^d \times [0,R]} \mu(x) \tag{453}$$

and

$$\left\| \tilde{f} \right\|_{\mathcal{B}_{\sigma_s}^{S^d \times [0,R]}} \leqslant 2 \frac{(1+R)^s}{s!} \|f\|_{\mathscr{F}^{s+1,1}(U)}. \tag{454}$$

Note that we did not include the smoothness requirement of theorem 4.2 in conjecture 8.2, since we have shown in proposition 6.4 that $\mathscr{F}_I^{s+1,1}$ is already sufficiently smooth.

It is also possible to write $\mathscr{F}_I^{s,1}(\mathcal{X})$ more as a Barron space. For this we use section 7 of [Caragea et al., 2020]. The authors discuss the space

$$\begin{aligned} \mathcal{B}_{\mathcal{F},s}(U) &= \left\{ f : U \to \mathbb{R} \;\middle|\; \exists F : \mathbb{R}^d \to \mathbb{C} : f(x) = \int_{\mathbb{R}^d} e^{i\langle x|\xi\rangle} |F(\xi)| d\xi, \int_{\mathbb{R}^d} (1 + \|\xi\|)^s F(\xi) d\xi < \infty \right\}, \\ \|f\|_{\mathcal{B}_{\mathcal{F},s}} &= \inf \left\{ \int_{\mathbb{R}^d} (1 + \|\xi\|)^s |F(\xi)| d\xi \;\middle|\; F : \mathbb{R}^d \to \mathbb{C} \text{ measurable}, f(x) = \int_{\mathbb{R}^d} e^{i\langle x|\xi\rangle} F(\xi) d\xi \right\}. \end{aligned} \tag{455}$$

This formulation is a different way of writing $\mathscr{F}_I^{s,1}(\mathcal{X})$. However, it helps us to write it as a Barron space. Namely, we can combine $F(\xi)d\xi$ into a single complex valued measure $\gamma \in rca(\mathbb{R}^d, \mathbb{C})$ such that

$$F(\xi) d\xi = d\gamma(\xi). \tag{456}$$

Then, we can split the complex measure γ into into a complex part and a real part

$$d\gamma(\xi) = e^{i\theta(\xi)} d\nu(\xi) \tag{457}$$

with $\theta : \mathbb{R}^d \to \mathbb{R}$ and $\nu \in rca(\mathbb{R}^d)$. Finally, for each pair (θ, ν) we can find a measure $\mu \in rca(\mathbb{R}^{d+1})$ such that

$$\int_{\mathbb{R}^{d+1}} e^{i(\langle x|\xi\rangle + b)} d\mu(\xi, b) = \int_{\mathbb{R}^d} e^{i\langle x|\xi\rangle} e^{i\theta(\xi)} d\nu(\xi). \tag{458}$$

Combining eq. (455) till eq. (458) and relabelling ξ as w gives

$$\begin{aligned} \mathcal{B}_{\mathscr{F},s}(U) &= \left\{ f : U \to \mathbb{R} \;\middle|\; \|f\|_{\mathscr{F}_I^{s,1}(U)} < \infty \right\}, \\ \|f\|_{\mathcal{B}_{\mathscr{F},s}} &= \inf \left\{ \int_{\mathbb{R}^{d_1}} (1 + \|w\|)^s d|\mu(w, b)| \;\middle|\; \mu \in rca(\mathbb{R}^{d+1}), f(x) = \int_{\mathbb{R}^{d+1}} e^{i(\langle x|w\rangle + b)} d\mu(w, b) \right\}. \end{aligned} \tag{459}$$

$\mathcal{B}_{\mathscr{F},s}$ can be seen as a version of $\mathcal{N}^{\Omega}_{\phi,W}$ where the activation function ϕ, the domain Ω and weight function W are given by

$$\begin{aligned}\phi &: x \mapsto e^{ix}\\ \Omega &= \mathbb{R}^{d+1}\\ W &: (w,b) \mapsto (1+\|w\|)^s\end{aligned}$$

respectfully. Note that this at most shows that

$$\mathscr{F}_I^{s,1}(U) \simeq \mathcal{B}_{\mathcal{F},s}(U) \subseteq \mathcal{B}_{\mathscr{F},s}(U) \tag{460}$$

for all $U \subseteq \mathbb{R}^d$. The reverse should be provable.

Conjecture 8.3. *Let $s \in \mathbb{N}$. For all $U \subseteq \mathbb{R}^d$*

$$\mathscr{F}_I^{s,1}(U) \simeq \mathcal{B}_{\mathcal{F},s}(U) \simeq \mathcal{B}_{\mathscr{F},s}(U). \tag{461}$$

In section 7 of [Caragea et al., 2020] the authors also prove some relationships between some Barron spaces and $\mathcal{B}_{\mathcal{F},s}$. These are shown in fig. 11. Recall that in section 6.1 we argued that $\mathscr{F}_I^{2,1}$ does not embed in $\mathcal{B}_\sigma^{\mathbb{S}^d \times [0,R]}$ for any $R \geqslant 0$, but that some local version of $\mathscr{F}_I^{2,1}$ may embed in $\mathcal{B}_\sigma^{\mathbb{S}^d \times [0,R]}$. If conjecture 8.3 holds, then the authors have shown that a similar statement is indeed true.

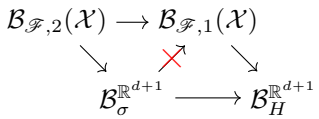


Figure 11: Representations of relations between some Barron spaces and $\mathcal{B}_{\mathcal{F}^{s,1}}$ as proven by the authors in [Caragea et al., 2020]. Arrows represent embeddings. The arrow with a red cross implies that no such embedding can exist.

5. Is it possible to construct an inverse of the operator K?

In [Parhi and Nowak, 2021] the authors discuss a similar result as theorem 4.2. They state that any function $f \in \mathcal{F}_{s+1}$ with

$$\mathcal{F}_s = \left\{ f : \mathbb{R}^d \to \mathbb{R} \;\middle|\; \operatorname*{ess\,sup}_{x\in\mathbb{R}^d} |f(x)|(1+\|x\|_{\ell^2})^{1-s}, R_s f \in rca(\mathbb{S}^d \times \mathbb{R}) \right\} \tag{462}$$

where R_s is a particular operator that we will discuss in a bit, can be represented as a (infinitely wide) shallow neural network using a higher order ReLU combined with a polynomial. More precisely, let P_m be the set of polynomials of order at most m, $p_i \in P_m$ be a basis of P_m and $q_i \in P_m^*$ be dual elements such that

$$\langle p_i | q_j \rangle_{L^2(\mathbb{R}^d)} = \int_{\mathbb{R}^d} p_i(x) q_j(x) dx = \begin{cases} 1 & i = j, \\ 0 & i \neq j. \end{cases} \tag{463}$$

For every $f \in \mathcal{F}_{s+1}$ there exists a measure $\mu \in rca(\mathbb{S}^d \times \mathbb{R})$ and a polynomial h of degree at most s such that for all $x \in \mathbb{R}^d$

$$f(x) = h(x) + \int_{\mathbb{S}^d \times \mathbb{R}} k_s(x, (w, b)) d\mu(w, b) \tag{464}$$

with

$$k_s(x, (w, b)) = \sigma_s(\langle x|w\rangle + b) - \sum_{i=1}^{m} p_i(x) \int_{\mathbb{R}^d} q_i(y)\sigma_s(\langle y|w\rangle + b)dy. \tag{465}$$

Note that the biggest difference between eq. (464) and eq. (236) is that eq. (464) uses the kernel k_s whereas eq. (236) only has $\sigma_s(\langle x|w\rangle + b)$. To prove their claim the authors use that

$$R_{s+1}h = 0 \tag{466}$$

and that

$$R_{s+1}\sigma_s(\langle \cdot|w\rangle - b) = \frac{\delta_{-(w,b)} + (-1)^s \delta_{(w,b)}}{2}. \tag{467}$$

This allows them to show that

$$R_{s+1} \int_{\mathbb{S}^d \times \mathbb{R}} k_s(\cdot, (w, b)) d\mu(w, b) = \mu. \tag{468}$$

Hence, it makes sense to call

$$R_{s+1}^{-1} : \mu \mapsto \int_{\mathbb{S}^d \times \mathbb{R}} k_s(\cdot, (w, b)) d\mu(w, b) \tag{469}$$

the right inverse of R_{s+1}. Recall that K_ϕ^Ω was given by

$$K_{\sigma_s}^\Omega : \mu \mapsto \int_\Omega \sigma_s(\langle \cdot|w\rangle + b) d\mu(w, b), \tag{470}$$

for the higher order ReLU σ_s. The similarity between eq. (469) and eq. (470) suggests that we can alter R_{s+1} to get a left inverse for K. How to do this is unclear to the author.

References

at Caltech, A. R. a. C. (2020). Neural-Swarm2: Planning and Control of Heterogeneous Multirotor Swarms using Learned Interactions. https://www.youtube.com/watch?v=Y02juH6BDxo

Bach, F. (2017). Breaking the Curse of Dimensionality with Convex Neural Networks. Journal of Machine Learning Rese 18(19), 1–53. http://jmlr.org/papers/v18/14-546.html

Barron, A. (1993). Universal approximation bounds for superpositions of a sigmoidal function. IEEE Transactions on Information Theory, 39(3), 930–945. https://doi.org/10.1109/18.256500

Batra, S., Huang, Z., Petrenko, A., Kumar, T., Molchanov, A., & Sukhatme, G. S. (2021). RL Quadrotor Swarms, CoRL 2021. Retrieved April 17, 2022, from https://sites.google.com/view/swarm-rl

Caragea, A., Petersen, P., & Voigtlaender, F. (2020). Neural network approximation and estimation of classifiers with classification boundary in a Barron class [arXiv: 2011.09363]. arXiv:2011.09363 [math, stat]. Retrieved April 22, 2021, from http://arxiv.org/abs/2011.09363

E, W., Ma, C., Wojtowytsch, S., & Wu, L. (2020). Towards a Mathematical Understanding of Neural Network-Based Machine Learning: What we know and what we don't [arXiv: 2009.10713]. arXiv:2009.10713 [cs, math, stat]. Retrieved April 15, 2021, from http://arxiv.org/abs/2009.10713

E, W., Ma, C., & Wu, L. (2021). The Barron Space and the Flow-induced Function Spaces for Neural Network Models [arXiv: 1906.08039]. arXiv:1906.08039 [cs, math, stat]. Retrieved May 13, 2021, from http://arxiv.org/abs/1906.08039

E, W., & Wojtowytsch, S. (2020a). Kolmogorov Width Decay and Poor Approximators in Machine Learning: Shallow Neural Networks, Random Feature Models and Neural Tangent Kernels [arXiv: 2005.10807]. arXiv:2005.10807 [cs, math, stat]. Retrieved June 10, 2021, from http://arxiv.org/abs/2005.10807

E, W., & Wojtowytsch, S. (2020b). Representation formulas and pointwise properties for Barron functions [arXiv: 2006.05982]. arXiv:2006.05982 [cs, math, stat]. Retrieved April 15, 2021, from http://arxiv.org/abs/2006.05982

Grafakos, L. (2014a). Classical Fourier Analysis (Vol. 249). Springer New York. https://doi.org/10.1007/978-1-4939-1194-3

Grafakos, L. (2014b). Modern Fourier Analysis (Vol. 250). Springer New York. https://doi.org/10.1007/978-1-4939-1230-8

Kan, X., Thomas, J., Teng, H., Tanner, H. G., Kumar, V., & Karydis, K. (2019). Analysis of Ground Effect for Small-Scale UAVs in Forward Flight. IEEE Robotics and Automation Letters, 4(4), 3860–3867. https://doi.org/10.1109/LRA.2019.2929993

Klusowski, J. M., & Barron, A. R. (2018). Risk Bounds for High-dimensional Ridge Function Combinations Including Neural Networks [arXiv: 1607.01434]. arXiv:1607.01434 [math, stat]. Retrieved April 19, 2021, from http://arxiv.org/abs/1607.01434

Kolyada, V. I. (1997). Estimates of Fourier transforms in Sobolev spaces. Studia Mathematica, 1(125), 67–74. Retrieved February 17, 2022, from https://www.infona.pl//resource/bwmeta1.element.bwnjournal-article-smv125i1p67bwm

Li, Z., Ma, C., & Wu, L. (2020). Complexity Measures for Neural Networks with General Activation Functions Using Path-based Norms [arXiv: 2009.06132]. arXiv:2009.06132 [cs, stat]. Retrieved June 24, 2021, from http://arxiv.org/abs/2009.06132

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). Foundations of machine learning. MIT Press.

O'Connell, M., Shi, G., Shi, X., & Chung, S.-J. (2021). Meta-Learning-Based Robust Adaptive Flight Control Under Uncertain Wind Conditions [arXiv: 2103.01932]. arXiv:2103.01932 [cs, eess]. Retrieved April 17, 2022, from http://arxiv.org/abs/2103.01932

Parhi, R., & Nowak, R. D. (2021). Banach Space Representer Theorems for Neural Networks and Ridge Splines. Journal of Machine Learning Research, 22(43), 1–40. http://jmlr.org/papers/v22/20-583.html

Rudin, W. (2006). Functional analysis (2. ed) [OCLC: 711823671]. Tata McGraw Hill Education.

Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press. https://doi.org/10.1017/CBO9781107298019

Shi, G., Shi, X., O'Connell, M., Yu, R., Azizzadenesheli, K., Anandkumar, A., Yue, Y., & Chung, S.-J. (2019). Neural Lander: Stable Drone Landing Control using Learned Dynamics [arXiv: 1811.08027]. 2019 International Conference on Robotics and Automation (ICRA), 9784–9790. https://doi.org/10.1109/ICRA.2019.8794351

Siegel, J. W., & Xu, J. (2021). Approximation Rates for Neural Networks with General Activation Functions [arXiv: 1904.02311]. arXiv:1904.02311 [cs, math]. Retrieved April 15, 2021, from http://arxiv.org/abs/1904.02311

Sobolev inequality [Page Version ID: 1059603306]. (2021). Retrieved February 19, 2022, from https://en.wikipedia.org/w/index.php?title=Sobolev_inequality&oldid=1059603306

Tao, T. (2009). 245C, Notes 4: Sobolev spaces. Retrieved February 18, 2022, from https://terrytao.wordpress.com/2009/04/30/245c-notes-4-sobolev-spaces/

Visintin, A. (2017). Notes on Sobolev Spaces. Retrieved February 19, 2022, from https://www.science.unitn.it/~visintin/Sobolev2017.pdf

Wolf, M. M. (2018). Mathematical Foundations of Supervised Learning – lecture notes. https://www-m5.ma.tum.de/foswiki/pub/M5/Allgemeines/MA4801_2016S/ML_notes_main.pdf