Embeddings between Barron spaces with higher order activation functions

Tjeerd Jan Heeringa[1,*], Len Spek[1], Felix Schwenninger[2], and Christoph Brune[1]

[1]*Mathematics of Imaging & AI, University of Twente, Enschede, The Netherlands*
[2]*Mathematics of Systems Theory, University of Twente, Enschede, The Netherlands*
[*]*Corresponding author: t.j.heeringa@utwente.nl*

2023-05-25

Abstract

The approximation properties of infinitely wide shallow neural networks heavily depend on the choice of the activation function. To understand this influence, we study embeddings between Barron spaces with different activation functions. These embeddings are proven by providing push-forward maps on the measures μ used to represent functions f. An activation function of particular interest is the rectified power unit (RePU) given by $\mathrm{RePU}_s(x) = \max(0, x)^s$. For many commonly used activation functions, the well-known Taylor remainder theorem can be used to construct a push-forward map, which allows us to prove the embedding of the associated Barron space into a Barron space with a RePU as activation function. Moreover, the Barron spaces associated with the RePU_s have a hierarchical structure similar to the Sobolev spaces H^m.

keywords: Neural networks, activation functions, Barron spaces, push-forward, rectified power unit

1 Introduction

For a given function $\sigma : \mathbb{R} \to \mathbb{R}$, called the activation function, and a set $\mathcal{X} \subseteq \mathbb{R}^d$, we consider the set of functions that can be written using an integral representation as

$$f(x) = \int_\Omega \sigma(\langle x|w \rangle + b) d\mu(w, b), \quad x \in \mathcal{X} \tag{1.1}$$

where μ is a radon measure on $\Omega \subseteq \mathbb{R}^{d+1}$. In machine learning, this represents an infinitely wide shallow neural network.

The choice of activation function is a crucial aspect in the design and training of neural networks, as it directly affects their expressive power, convergence rate, and generalization performance. The search for new activation functions that can better capture the underlying structure and patterns of the data, while avoiding common issues such as vanishing gradients and overfitting, is ongoing. The most popular choices for the activation function are the sigmoidal functions[1] and the $\mathrm{ReLU}(x) := \max(0, x)$. Although these activation functions work in many instances, they are not the best for all instances. For example, [Siegel and Xu, 2021] showed that in certain instances taking a higher-order version of the ReLU given by $\mathrm{RePU}_s(x) := \max(0, x)^s$ yields improved approximation properties compared to using ReLU. Similarly, [Hendrycks and Gimpel, 2020;Ramachandran et al., 2023;Misra, 2020] showed that taking smoothed versions of ReLU like GeLU, Swish or Mish as activation functions also accomplished this. To understand what these changes to the activation function do and why they lead to better approximation properties,

---

[1]Sigmoidal functions are monotonic increasing functions which go to constants at plus and minus infinity.

we want to study how the function spaces associated with neural networks change when the activation function is changed.

Changing the activation function means that the natural norm of the associated vector space gets changed too. The natural norm of the space for deep neural networks is unknown. For shallow neural networks, the Barron norm gives the natural norm for an infinitely wide neural network with activation function σ. For sigmoidal activation functions, this is given by

$$\|f\|_{\mathcal{B}_{\sigma}} = \inf \int_{\Omega} (1 + \|w\|_{\ell^1} + |b|)d|\mu|(w, b) \tag{1.2}$$

where the infimum is taken over all measures μ such that (1.1) holds for all $x \in \mathcal{X}$. A vector space with this norm is called a *Barron space* (associated with the activation function σ)[E and Wojtowytsch, 2022b]. Hence, to understand what effect a change to the activation function has, we should determine what effect this change has on the underlying Barron space.

1.1   Related work

The Barron spaces were introduced with the motivation to create a "reasonably simple and transparent framework for machine learning"[Page 2 of E and Wojtowytsch, 2022b]. Initially, they were introduced only for the ReLU and sigmoidal activation functions. In [Bartolucci et al., 2023], Barron spaces were shown to be *reproducing kernel Banach spaces* (RKBS), a Banach space analogue to *reproducing kernel Hilbert spaces* (RKHS). This was used to determine what form the Barron norm should have for the *rectified power unit* (RePU), but can easily be extended to determine what a natural norm would be for any activation function. It was proven that Barron functions have bounded point evaluations [Bartolucci et al., 2023; Spek et al., 2023], Barron functions can be approximated in L^p with rate $O(m^{-1/p})$[E. and Wojtowytsch, 2022a], Barron spaces have a representer theorem[Parhi and Nowak, 2021] and more. Some of these results hold for particular activation functions (mostly ReLU), whereas others hold for more general classes of activation functions.

In order to extend results for the Barron space with ReLU as the activation function to more general activation functions, a relation between ReLU and a large class of activation functions was established in [Li et al., 2020]. They determined that any activation function ϕ that satisfies

$$\int_{\mathbb{R}} \left|D^2 \phi(x)\right|(1 + |x|)dx < \infty \tag{1.3}$$

can be approximated up to arbitrary precision in L^{∞} by a finite linear combination of ReLUs. This covers, among others, the sigmoidal activation functions. Their result does not apply directly to the infinite width setting of the Barron spaces, but their strategy allows for an extension to the infinite width setting.

In [Caragea et al., 2020], the relations between the Barron spaces and the related spectral Barron spaces were discussed. For $s \in \mathbb{N}$ the spectral Barron spaces have norm

$$\|f\|_{\mathcal{B}_{\mathscr{F},s}} = \inf \int_{\mathbb{R}^d} (1 + \|\xi\|_{\ell^1})^s \left|\hat{f}_e(\xi)\right| d\xi \tag{1.4}$$

where the infimum is taken over all extensions $f_e \in L^1(\mathbb{R}^d)$ of f. They can be seen as Barron spaces with a cosine as the activation function. It was shown that these spaces are closely related to but distinct from the Barron spaces with ReLU as the activation function. In particular, $s \geq 2$ needs to hold to have an embedding into the Barron space with ReLU as the activation function.

In [Siegel and Xu, 2020], it was shown that functions $f \in \mathcal{B}_{\mathscr{F},s+1}$ can be approximated in H^s with rate $O(m^{-0.5})$ when using activation functions $\sigma \in W_{loc}^{s,\infty}$, with which a finite linear combination can be formed that decays sufficiently fast. Their work does not provide an embedding between the respective spaces.

In practice, there are many more activation functions being used. ReLU6 and leaky ReLU (LReLU) are linear combinations of ReLUs [A. G. Howard et al., 2017;Maas, 2013]. Tanh is a convolution of ArcTan

with a specific kernel. SoftPlus and ReLU are derivatives of Sigmoid and squared ReLU respectively [Glorot et al., 2011]. HardSwish, SILU/Swish-1, GeLU and the growing cosine unit are ReLU6, Sigmoid, Gaussian normal CDF function, and cosine respectively multiplied by their input [A. Howard et al., 2019;Ramachandran et al., 2023;Hendrycks and Gimpel, 2020;Noel et al., 2021]. What these and other changes to the activation do to the associated Barron space is unknown.

1.2   Our contribution

In this work, we show that many of these changes to the activation function lead to an embedding between the respective Barron spaces. The main idea is that we explicitly construct a push-forward map Θ for two given activation functions σ and ϕ so that

$$f(x) = \int_\Omega \sigma(\langle x|w\rangle + b)d\mu(w,b) = \int_\Omega \phi(\langle x|w\rangle + b)d\Theta_\#\mu(w,b), \quad x \in \mathcal{X}. \tag{1.5}$$

When we find a map Θ, we can use it to show that the Barron norm $\|f\|_{\mathcal{B}_\phi}$ is finite and determine the constant of embedding by using the relation Θ induces between $\Theta_\#\mu$ and μ. For example, for two Lipschitz continuous activation functions σ and ϕ we need Θ to be such that

$$\|f\|_{\mathcal{B}_\phi} \leq \int_\Omega 1 + \|w\|_{\ell^1} + |b|d|\Theta_\#\mu|(w,b) \lesssim \int_\Omega 1 + \|w\|_{\ell^1} + |b|d|\mu|(w,b) \tag{1.6}$$

for all μ satisfying (1.1) to get an embedding. The embeddings can be grouped into those in which one of the two activation functions is the RePU_s for some $s \in \mathbb{N}$ and those in which neither activation function is a RePU_s. The former is discussed in Section 2, whilst the latter is discussed in Section 3. Additionally, in Section 4 we show how the push-forward strategy can be used to provide embeddings from non-Barron spaces into Barron spaces by proving the embedding of the spectral Barron spaces into the Barron spaces with RePU activation. The proven embeddings are summarized in Theorem 1.

Theorem 1. *Let $s \in \mathbb{N}$. If ψ and ϕ are Lipschitz activation functions such that*

$$\phi(x) = \int_{\mathbb{R}^2} \psi(xw + b)d\gamma(w,b) \tag{1.7}$$

for all $x \in \mathbb{R}$ and for some measure $\gamma \in \mathcal{M}(\mathbb{R}^2)$ satisfying

$$\int_{\mathbb{R}^2}(1 + |w| + |b|)d|\gamma|(w,b) < \infty, \tag{1.8}$$

then

1. *$\mathcal{B}_\phi \hookrightarrow \mathcal{B}_\psi$,*

2. *$\mathcal{B}_\psi \hookrightarrow \mathcal{B}_{\mathrm{RePU}_1}$ whenever $\psi \in C^1(\mathbb{R})$ with $D^2\psi \in L^1(\mathbb{R})$,*

3. *$\mathcal{B}_\psi \hookrightarrow \mathcal{B}_{\mathrm{RePU}_s}$ whenever $\psi \in C^s(\mathbb{R})$ with $D^{s+1}\psi \in L^1(\mathbb{R})$ and Ω is bounded,*

4. *$\mathcal{B}_{\mathrm{RePU}_s} \hookrightarrow \mathcal{B}_{\mathrm{RePU}_t}$ for $t \in \mathbb{N}$ with $t \leq s$.*

Moreover, $\mathcal{B}_{\mathscr{F},s+1} \hookrightarrow \mathcal{B}_{\mathrm{RePU}_s}$.

Observe that the form of the integral in (1.7) is similar to that of the integrals in (1.6). This means we can interpret the embedding in point 1) of Theorem 1 as follows: If we have a shallow neural network with ϕ as activation function representing the function f, then we can replace each neuron in the hidden layer by a (possibly infinite) number of neurons with ψ as activation function. γ describes how the weights and biases of the neurons in the network should be adjusted. After the replacement, the network will still represent f. This interpretation is visualized in Figure 1.
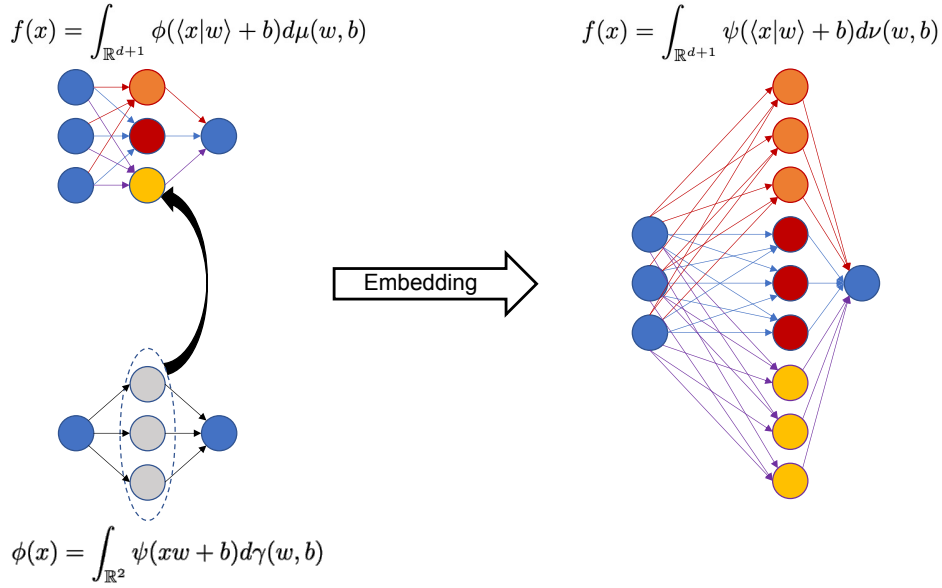
$$f(x) = \int_{\mathbb{R}^{d+1}} \phi(\langle x|w \rangle + b) d\mu(w, b)$$

$$f(x) = \int_{\mathbb{R}^{d+1}} \psi(\langle x|w \rangle + b) d\nu(w, b)$$

Embedding

$$\phi(x) = \int_{\mathbb{R}^2} \psi(xw + b) d\gamma(w, b)$$

Figure 1: Each circle represents a neuron, and arrows represent connections between neurons. On the left, a network with ϕ as activation function representing f is shown. The activation function ϕ can be represented using a shallow neural network with 3 neurons in the hidden layer and activation function ψ. On the right, a network with ψ as activation function representing f is shown. The network representing ϕ is used to construct the network on the right from that on the top left. Colors have been added to track which neuron on the right corresponds to which on the top left.

1.3 Notation

We use the following notation conventions throughout this paper.

Denote with $\mathbb{N}$ the natural numbers without zero. The weak derivative of a function f is denoted by Df and the (classical) derivative by ∂f. When f is multivariate, we use multi-indices to denote the partial derivatives. Radon measures are regular signed Borel measures with bounded total variation. The space of Radon measures $\mathcal{M}(\Omega)$ on a locally compact Hausdorff space Ω is the continuous dual of the continuous functions vanishing at infinity, $C_0(\Omega)^* = \mathcal{M}(\Omega)$. The total variation measure of a measure $\mu \in \mathcal{M}(\Omega)$ is denoted by $|\mu|$. The Dirac measure is given by

$$\delta_w(A) = \begin{cases} 1 & w \in A \\ 0 & w \notin A \end{cases} \tag{1.9}$$

for Borel sets $A \subseteq \Omega$. For a map Θ defined by

$$\Theta : X \to Y, \quad x \mapsto \Theta(x), \tag{1.10}$$

we call the measure $\nu := \Theta_{\#}\mu$ the push-forward of μ along the map Θ such that

$$\int_Y f(y) d\nu(y) = \int_X f(\Theta(x)) d\mu(x), \tag{1.11}$$

for all ν-measurable functions f. A normed vector space A embeds into another normed vector space B if and only if $A \subseteq B$ and $\|f\|_B \lesssim \|f\|_A$ for all $f \in A$, where $\lesssim$ means that the inequality holds up to a constant $C > 0$ independent of f. If $f \in L^1(\mathbb{R}^d)$, then

$$\hat{f}(\xi) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i\langle x|\xi \rangle} f(x) d\xi, \quad \xi \in \mathbb{R}^d \tag{1.12}$$

denotes the Fourier transform of f. As is common, we also use $\hat{f}$ if the Fourier transform exists in a generalized way.

2 Embeddings between Barron spaces involving RePU

In this section, we start by defining the Barron spaces in Section 2.1. We proceed by showing that Barron spaces with an activation function for which the weak derivative is in $L^1(\mathbb{R})$ embed into the Barron space with ReLU as activation function. After assuming Ω is bounded, we extend this to: Barron spaces with an activation function $\sigma \in C^s(\mathbb{R})$ for which $D^{s+1}\sigma \in L^1(\mathbb{R})$ embed in a Barron spaces with a RePU_s. Next to that, we show that the Barron spaces with RePU as the activation function have a hierarchical structure. The former two we do in Section 2.2 and the latter in Section 2.3.

2.1 Barron spaces

The definition of the Barron spaces that we will be using is an adaption of Definition A.2 of [E and Wojtowytsch, 2021]. We use signed Radon measures instead of probability measures and have defined a natural norm for when the activation function is a RePU.

Fix $d \in \mathbb{N}$. Let $\mathcal{X} = [-1,1]^d$ and $\Omega \subseteq \mathbb{R}^{d+1}$. When we write $(w,b) \in \Omega$, we mean that $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$. We use the ℓ^1 norm for $(w,b) \in \Omega$ and the ℓ^∞ norm for $x \in \mathcal{X}$ so that $\|(w,b)\|_{\ell^1} = \|w\|_{\ell^1} + |b|$ and $|\langle x|w\rangle| \le \|w\|_{\ell^1}$. We call $\sigma : \mathbb{R} \to \mathbb{R}$ an activation function when it is Lipschitz continuous or $\sigma(x) = \text{RePU}_s(x) := \max(0,x)^s$ for some $s \in \mathbb{N}$. Consider functions $f : \mathcal{X} \to \mathbb{R}$ given by

$$f(x) = \int_\Omega \sigma(\langle x|w\rangle + b)d\mu(w,b), \quad x \in \mathcal{X}. \tag{2.1}$$

Since several distinct measures μ can describe the same function f, we group them into

$$\mathbb{G}_{\sigma,f} = \left\{ \mu \in \mathcal{M}(\Omega) \;\middle|\; \forall x \in \mathcal{X} : \; f(x) = \int_\Omega \sigma(\langle x|w\rangle + b)d\mu(w,b) \right\}. \tag{2.2}$$

The Barron norm is given by

$$\|f\|_{\mathcal{B}_\sigma} = \inf_{\mu \in \mathbb{G}_{\sigma,f}} \int_\Omega (1 + \|w\|_{\ell^1} + |b|)d|\mu|(w,b), \tag{2.3}$$

unless $\sigma = \text{RePU}_s$ for some $s \in \mathbb{N}$ in which case

$$\|f\|_{\mathcal{B}_\sigma} = \inf_{\mu \in \mathbb{G}_{\sigma,f}} \int_\Omega (\|w\|_{\ell^1} + |b|)^s d|\mu|(w,b). \tag{2.4}$$

The Barron space for a given activation function σ is given by

$$\mathcal{B}_\sigma = \left\{ f : \mathcal{X} \to \mathbb{R} \;\middle|\; \|f\|_{\mathcal{B}_\sigma} < \infty \right\}. \tag{2.5}$$

2.2 Absolutely continuous activation functions

To show that the Barron spaces with an activation function $\sigma \in C^s(\mathbb{R})$ for which $D^{s+1}\sigma \in L^1(\mathbb{R})$ embed into a Barron space with a RePU as activation function, we first prove some technical lemmas concerning RePU_s, which are important for the proofs of the embeddings. These lemmas will be reused in Section 4.

Recall that the rectified power unit is given by

$$\text{RePU}_s(x) = \max(0,x)^s \tag{2.6}$$

for $s \in \mathbb{N}$. The embeddings proven in this section rely on the tie between RePU_s and the Taylor remainder theorem to construct the push-forward map. The integral form of the Taylor remainder theorem states

that a function $\phi \in C^s(\mathbb{R})$ of which $\partial^s \phi$ is absolutely continuous on the closed interval $[a, b]$, can be written as

$$\phi(x) = \underbrace{\sum_{k=0}^{s} \frac{\partial^k \phi(y)}{k!} (x-y)^k}_{\text{series}} + \underbrace{\int_y^x \frac{D^{s+1}\phi(t)}{s!} (x-t)^s dt}_{\text{remainder}} \tag{2.7}$$

for all $x, y \in [a, b]$. This well-known theorem follows straightforwardly from several applications of integration by parts. For fixed y, both the series and remainder parts can be written in the form of (1.1) using a suitably chosen measure. To prove this, we use the following two lemmas. The first lemma deals with the series part. This lemma is similar to Theorem 2 in [Chen et al., 2022]. However, the proof for the full statement in [Chen et al., 2022] is contained in a currently unpublished paper. Hence, for completeness, we have provided this proof.

Lemma 2.1. *Let $s, d \in \mathbb{N}$. There exist $p := \binom{s+d}{d}$ pairs $(w_i, b_i) \in \mathbb{R}^{d+1}$ with $i \in \{1, \dots, p\}$ such that the set of polynomials $\Theta_s := \left\{ (\langle x|w_1 \rangle + b_1)^s, (\langle x|w_2 \rangle + b_2)^s, \dots, (\langle x|w_p \rangle + b_p)^s \right\}$ forms a basis for the space of polynomials in d variables with degree at most s.*

Proof. There are p multi-indices with total degree at most s. Let the sequence $\{\alpha_n\}_{n=1}^p$ be the set with these multi-indices in inverse lexicographical order. The statement holds if we can choose the pairs $(w_i, b_i) \in \mathbb{R}^{d+1}$ such that there exists an invertible matrix W which satisfies

$$WX = P_s \tag{2.8}$$

where $X = \left(x^{\alpha_1}, \dots, x^{\alpha_p} \right)^\mathsf{T}$ and $P_s = \left((\langle x|w_1 \rangle + b_1)^s, \dots, (\langle x|w_p \rangle + b_p)^s \right)^\mathsf{T}$. We will construct W and use the theory of generalized Vandermonde matrices to show that it is invertible [Gantmacher, 2009].

Observe that for $(w, b) \in \Omega$ and $x \in \mathcal{X}$ we have by simple combinatorics that

$$\begin{aligned}
(\langle x|w \rangle + b)^s &= (\langle (x, 1)|(w, b) \rangle)^s \\
&= \sum_{|\beta|=s} \binom{s}{\beta} (w, b)^\beta (x, 1)^\beta \\
&= \sum_{|\gamma| \le s} \binom{s}{|\gamma|} \binom{|\gamma|}{\gamma} w^\gamma b^{s-|\gamma|} x^\gamma \\
&= \sum_{n=1}^p \binom{s}{|\alpha_n|} \binom{|\alpha_n|}{\alpha_n} w^{\alpha_n} b^{s-|\alpha_n|} x^{\alpha_n},
\end{aligned} \tag{2.9}$$

where β and γ are multi-indices of length $d+1$ and length d respectively. This shows that a matrix W with elements of the form $W_{ij} = \binom{s}{|\alpha_j|} \binom{|\alpha_j|}{\alpha_j} w_i^{\alpha_j} b_i^{s-|\alpha_j|}$ satisfies (2.8). What remains is choosing each (w_i, b_i) such that W is invertible.

If $\tilde{W}$ is a matrix with elements $w_i^{\alpha_j} b_i^{s-|\alpha_j|}$ and D a diagonal matrix with entries $D_{jj} = \binom{s}{|\alpha_j|} \binom{|\alpha_j|}{\alpha_j}$, then

$$det(W) = det(D) det(\tilde{W}). \tag{2.10}$$

Clearly, $det(D) > 0$. If we take $b_i = 1$, $1 < w_{1,1} < w_{2,1} < \dots w_{p,1} < \infty$, and $w_{i,k} = w_{i,1}^{1+(k-1)\sqrt{\text{prime}(k)}}$, where $\text{prime}(k)$ is the kth prime number, for all $i \in \{1, \dots p\}$, then each element of $\tilde{W}$ is of the form

$$\tilde{W}_{ij} = w_{i,1}^{|\alpha_j| + \sum_{k=1}^{d} (k-1)\sqrt{\text{prime}(k)} \alpha_{j,k}}. \tag{2.11}$$

The bases are fixed columnwise, but strictly increasing rowwise. The exponents are fixed rowwise, but distinct columnwise. Let $\tilde{\tilde{W}}$ be $\tilde{W}$ with its columns reordered such that the exponents are in increasing order. By construction, $\tilde{\tilde{W}}$ is a generalized Vandermonde matrix. These matrices have a non-zero determinant [page 99 of Gantmacher, 2009]. Reordering the columns at most switches the sign of the determinant. Hence, $\tilde{\tilde{W}}$ is invertible and thus W is too.   Q.E.D.

Using this lemma, we are now able to prove the following equality.

Lemma 2.2. *Let $p : \mathbb{R}^d \to \mathbb{R}$ be a polynomial of degree less or equal to $s \in \mathbb{N}$. Then there exists a measure $\nu \in \mathcal{M}(\mathbb{R}^{d+1})$ such that*

$$p(x) = \int_{\mathbb{R}^{d+1}} \mathrm{RePU}_s(\langle x | w \rangle + b) d\nu(w, b), \qquad x \in \mathbb{R}^d. \tag{2.12}$$

Proof. We can use Lemma 2.1 to write the polynomial p as a linear combination of the basis functions in Θ_s,

$$p(x) = \sum_{i=1}^{p} \kappa_i (\langle x | w_i \rangle + b_i)^s \tag{2.13}$$

where $\kappa_i \in \mathbb{R}$. Combined with the identity

$$z^s = \mathrm{RePU}_s(z) + (-1)^{s-1} \mathrm{RePU}_s(-z) \tag{2.14}$$

for all $z \in \mathbb{R}$, we can conclude that the measure $\nu = \nu_1 + \nu_2$ defined using

$$
\begin{aligned}
\nu_1 &= \sum_{i=1}^{p} \kappa_i \delta_{(w_i, b_i)} \\
\nu_2 &= \sum_{i=1}^{p} (-1)^{s-1} \kappa_i \delta_{(-w_i, -b_i)}
\end{aligned}
\tag{2.15}
$$

satisfies (2.12). $\hfill Q.E.D.$

There are several things to note regarding (the proofs of) Lemma 2.1 and Lemma 2.2. First, we chose w_i with $\|w_i\|_{\ell^1} \leq d|w_{i,1}|^{(d-1)\sqrt{\mathrm{prime}(d)}}$ and $w_{i,1} > 1$. This upper bound scales exponentially with dimension. Different choices for $w_{p,k}$ are available, like choosing $w_{i,k} = w_{i,1}^{\frac{(k-1)\sqrt{\mathrm{prime}(k)}}{d\sqrt{\mathrm{prime}(d)}}}$ with $w_{i,1}$ sufficiently small gives $\|w_i\|_{\ell^1} \leq 2d$. Second, both proofs are proven for $d \in \mathbb{N}$, whereas activation functions are univariate functions. We will use the higher dimensional case when discussing the spectral Barron spaces in Section 4. Last, an argument for deeper neural networks is that neural network with ReLU as its activation function require several layers to approximate these higher order monomials well [e.g. DeVore et al., 2021]. Our results suggest that instead of increasing the number of layers, we could increase the order of the RePU_s.

Lemma 2.2 can be used to write the $(x - t)^s$ in the remainder part of (2.7) as a linear combination of RePU_s's. However, this does not allow us to write the remainder part using a single measure $\nu \in \mathcal{M}(\mathbb{R}^2)$ for all $x \in \mathbb{R}$, because the integral bounds depend on x. In the second lemma, we deal with this by extending the domain of integration.

Lemma 2.3. *Let $s \in \mathbb{N}$ and $c > 0$. When $f \in L^1([-c, c])$, we have*

$$\int_0^z f(u)(z-u)^s du = \int_0^c f(u) \mathrm{RePU}_s(z-u) + (-1)^{s-1} f(-u) \mathrm{RePU}_s(-z-u) du \tag{2.16}$$

for all $z \in [-c, c]$.

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

$$\int_0^z (z-u)^s f(u) du = \begin{cases} (-1)^{s-1} \int_0^c (-z-u)^s \mathrm{Step}(-z-u) f(-u) du & -c \leq z \leq 0 \\ \int_0^c (z-u)^s \mathrm{Step}(z-u) f(u) du & 0 \leq z \leq c \end{cases} \tag{2.17}$$

where Step is the Heaviside step function. Note that the term $(-1)^{s-1}$ restores the sign for even s. Since both representations are zero in the domain of the other, we can add them to obtain

$$\int_0^z (z-u)^s f(u) du = \int_0^c (z-u)^s \mathrm{Step}(z-u) f(u) + (-1)^{s-1} (-z-u)^s \mathrm{Step}(-z-u) f(-u) du. \tag{2.18}$$

Observe that

$$
\begin{aligned}
(z-u)^s \operatorname{Step}(z-u) &= \operatorname{RePU}_s(z-u), \\
(-z-u)^s \operatorname{Step}(-z-u) &= \operatorname{RePU}_s(-z-u).
\end{aligned}
\tag{2.19}
$$

Substitution of (2.19) into (2.18) gives (2.16). $\hspace{6cm}$ $Q.E.D.$

From Lemma 2.3 and Lemma 2.2 it follows that both the series part and the remainder part of a function satisfying the Taylor remainder theorem can be written in terms of RePU_s. Hence, it may seem logical that an embedding of $\mathcal{B}_\phi$ into $\mathcal{B}_{\operatorname{RePU}_s}$ exists if ϕ satisfies the requirements for the Taylor remainder theorem. Without additional assumptions, this does not hold for all $s \in \mathbb{N}$. This is due to the different exponent in the norm for RePU_s for $s > 1$ compared to the exponent in the Barron norm for ϕ. We discuss this later in this section in more detail. For $s = 1$, this suggested embedding exists without additional assumptions. This is shown in the following proposition.

Proposition 2.1. *If $\psi \in C^1(\mathbb{R})$ such that $D^2\psi \in L^1(\mathbb{R})$, then we have $\mathcal{B}_\psi \hookrightarrow \mathcal{B}_{\operatorname{ReLU}}$ with*

$$
\|f\|_{\mathcal{B}_{\operatorname{ReLU}}} \leq \gamma(\phi)\|f\|_{\mathcal{B}_\psi}
\tag{2.20}
$$

for all $f \in \mathcal{B}_\psi$, where

$$
\gamma(\phi) := \inf_{y \in \mathbb{R}} \left(|\psi(y)| + 2|\partial\psi(y)| + 2(1+|y|) \int_{\mathbb{R}} |D^2\psi(z)| dz \right).
\tag{2.21}
$$

Proof. From the triangle inequality, it follows immediately that

$$
|\langle x|w \rangle + b - y| \leq \|w\|_{\ell^1} + |b| + |y| := \theta_{w,b,y}
\tag{2.22}
$$

for all $(x, y, w, b) \in [-1, 1]^d \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$. From the Taylor remainder theorem, it follows that for given $(w, b) \in \mathbb{R}^{d+1}$ and $y \in \mathbb{R}$

$$
\phi(\langle x|w \rangle + b) = \phi(y) + \partial^1\phi(y)(\langle x|w \rangle + b) + \int_y^{\langle x|w \rangle + b} D^2\phi(t)(\langle x|w \rangle + b) - t)dt
\tag{2.23}
$$

for all $x \in \mathcal{X}$. After the change of coordinate $v = t - y$, this becomes

$$
\phi(\langle x|w \rangle + b) = \phi(y) + \partial^1\phi(y)(\langle x|w \rangle + b) + \int_0^{\langle x|w \rangle + b - y} D^2\phi(v+y)(\langle x|w \rangle + b) - v - y)dv.
\tag{2.24}
$$

From Lemma 2.3, it follows that (2.24) is equivalent to

$$
\begin{aligned}
\phi(\langle x|w \rangle + b) = \phi(y) + \partial^1\phi(y)(\langle x|w \rangle + b) + \int_0^{\theta_{w,b,y}} & D^2\phi(v+y) \operatorname{ReLU}(\langle x|w \rangle + b - v - y) \\
& + D^2\phi(-v+y) \operatorname{ReLU}(\langle x|-w \rangle - b - v + y)dv
\end{aligned}
\tag{2.25}
$$

for all $x \in \mathcal{X}$. After the change of coordinate $\theta_{w,b,y} u = v$ and using (2.14), this becomes

$$
\begin{aligned}
\phi(\langle x|w \rangle + b) = \phi(y) \operatorname{ReLU}(1) + \partial^1\phi(y)\Big(& \operatorname{ReLU}(\langle x|w \rangle + b) - \operatorname{ReLU}(\langle x|-w \rangle - b) \Big) \\
& + \int_0^1 \theta_{w,b,y} D^2\phi(\theta_{w,b,y} u + y) \operatorname{ReLU}(\langle x|w \rangle + b - \theta_{w,b,y} u - y) \\
& + \int_0^1 \theta_{w,b,y} D^2\phi(-\theta_{w,b,y} u + y) \operatorname{ReLU}(\langle x|-w \rangle - b - \theta_{w,b,y} u + y)du.
\end{aligned}
\tag{2.26}
$$

Let $\mu \in \mathbb{G}_{\phi,f}$ for $f \in \mathcal{B}_\phi$. Observe that using (2.26) as a substitution we get

$$
f(x) = \int_{\mathbb{R}^{d+1}} \phi(\langle x|w \rangle + b)d\mu(w,b) = \int_{\mathbb{R}^{d+1}} \operatorname{ReLU}(\langle x|w \rangle + b)d\nu(w,b),
\tag{2.27}
$$

where the measure $\nu = \sum_{i=1}^{5} \nu_i$ is the sum of measures formed from the measures

$$
\begin{aligned}
\nu_1 &= \phi(y)\mu(\Omega)\delta_{(0,1)} \\
\nu_2 &= \partial^1 \phi(y)\mu \\
\nu_3 &= -\partial^1 \phi(y)\Theta^0_{\#}\mu \\
\nu_4(A) &= \int_A \int_0^1 \theta_{w,b,y} D^2 \phi(\theta_{w,b,y}u + y) d\Theta^1_{\#}(\mu \otimes \lambda)((w,b),u) \\
\nu_5(A) &= \int_A \int_0^1 \theta_{w,b,y} D^2 \phi(-\theta_{w,b,y}u + y) d\Theta^2_{\#}(\mu \otimes \lambda)((w,b),u)
\end{aligned}
\tag{2.28}
$$

for the Borel sets $A \subseteq \Omega$ and using the push-forward maps

$$
\begin{aligned}
\Theta^0 &: \mathbb{R}^{d+} \to \mathbb{R}^{d+1}, \ (w,b) \to (-w,-b) \\
\Theta^1 &: \mathbb{R}^{d+1} \times [0,1] \to \mathbb{R}^{d+1}, \ ((w,b),u) \to (w,b - \theta_{w,b,y}u - y) \\
\Theta^2 &: \mathbb{R}^{d+1} \times [0,1] \to \mathbb{R}^{d+1}, \ ((w,b),u) \to (-w,-b - \theta_{w,b,y}u + y)
\end{aligned}
\tag{2.29}
$$

where λ is the Lebesgue measure on $[0,1]$. Hence, $\nu \in \mathbb{G}_{\mathrm{ReLU},f}$. Furthermore, for each ν_i we have

$$
\int_{\mathbb{R}^{d+1}} (\|w\|_{\ell^1} + |b|) d|\nu_1|(w,b) \leq |\phi(y)||\mu|(\Omega) \leq |\phi(y)| \int_{\mathbb{R}^{d+1}} (1 + \|w\|_{\ell^1} + |b|) d|\mu|(w,b)
\tag{2.30}
$$

$$
\int_{\mathbb{R}^{d+1}} (\|w\|_{\ell^1} + |b|) d|\nu_2|(w,b) \leq |\partial\phi(y)||\mu|(\Omega) \leq |\partial\phi(y)| \int_{\mathbb{R}^{d+1}} (1 + \|w\|_{\ell^1} + |b|) d|\mu|(w,b)
\tag{2.31}
$$

$$
\int_{\mathbb{R}^{d+1}} (\|w\|_{\ell^1} + |b|) d|\nu_3|(w,b) \leq |\partial\phi(y)||\Theta^0_{\#}\mu|(\Omega) \leq |\partial\phi(y)| \int_{\mathbb{R}^{d+1}} (1 + \|w\|_{\ell^1} + |b|) d|\mu|(w,b)
\tag{2.32}
$$

and

$$
\begin{aligned}
& \int_{\mathbb{R}^{d+1}} (\|w\|_{\ell^1} + |b|) d(|\nu_4| + |\nu_5|)(w,b) \\
& \leq \int_{\mathbb{R}^{d+2}} \int_{-1}^1 \theta_{w,b,y} |D^2\phi(\theta_{w,b,y}u + y)| (1 + \|w\|_{\ell^1} + |b| + \theta_{w,b,y}|u| + |y|) du d|\mu|(w,b) \\
& \leq \sup_{(w,b)\in\mathbb{R}^{d+1}} \int_{-1}^1 \theta_{w,b,y} |D^2\phi(\theta_{w,b,y}u + y)| (1 + |u|)(1 + |y|) du \int_{\mathbb{R}^{d+1}} (1 + \|w\|_{\ell^1} + |b|) d|\mu|(w,b) \\
& \leq 2(1 + |y|) \sup_{(w,b)\in\mathbb{R}^{d+1}} \int_{-\theta_{w,b,y}+y}^{\theta_{w,b,y}+y} |D^2\phi(z)| dz \int_{\mathbb{R}^{d+1}} (1 + \|w\|_{\ell^1} + |b|) d|\mu|(w,b) \\
& = 2(1 + |y|) \int_{\mathbb{R}} |D^2\phi(z)| dz \int_{\mathbb{R}^{d+1}} (1 + \|w\|_{\ell^1} + |b|) d|\mu|(w,b)
\end{aligned}
\tag{2.33}
$$

where we used the change of coordinates $z = \theta_{w,b,y}u + y$. This means that by the triangle inequality

$$
\begin{aligned}
\|f\|_{\mathcal{B}_{\mathrm{ReLU}}} & \leq \int_{\mathbb{R}^{d+1}} (\|w\|_{\ell^1} + |b|) d|\nu|(w,b) \\
& \leq \sum_{i=1}^5 \int_{\mathbb{R}^{d+1}} (\|w\|_{\ell^1} + |b|) d|\nu_i|(w,b) \\
& = \left(|\phi(y)| + 2|\partial^1\phi(y)| + 2(1 + |y|) \int_{\mathbb{R}} |D^2\phi(z)| dz \right) \int_{\mathbb{R}^{d+1}} (1 + \|w\|_{\ell^1} + |b|) d|\mu|(w,b).
\end{aligned}
\tag{2.34}
$$

Taking the infimum over $\pi \in \mathbb{G}_{\phi,f}$ and $y \in \mathbb{R}$ gives (2.20). $Q.E.D.$

Observe that (2.21) is very similar to the definition of γ used in Theorem 1 of [Li et al., 2020],

$$
\gamma(\phi) = \inf_{y\in\mathbb{R}} \left(|\phi(y)| + (|y| + 2)|\partial\phi| + \int_{\mathbb{R}} |D^2\phi(t)|(1 + |t|)dt \right).
\tag{2.35}
$$

The change to our version of γ means that is sufficient for the function $D^2\phi(x)$ to go like $(1 + |x|)^{-(1+\epsilon)}$ instead of like $(1 + |x|)^{-(2+\epsilon)}$ for some $\epsilon > 0$ and for large values of x. Hence, Proposition 2.1 is satisfied for more activation functions than Theorem 1 of [Li et al., 2020].

Proposition 2.1 does not cover piecewise smooth activation functions, of which there are many. A slight alteration to Proposition 2.1 allows us to also cover activation functions that are smooth everywhere except at the origin. This can be found in Appendix A.1.

The right-hand side of (2.34) can be bounded so that

$$\|f\|_{\mathcal{B}_{\mathrm{ReLU}}} \lesssim \left(\|\phi\|_{C^1(\mathbb{R})} + \|D^2\phi\|_{L^1(\mathbb{R})} \right) \int_\Omega (1 + \|w\|_{\ell^1} + |b|) d|\mu|(w,b) \tag{2.36}$$

When we repeat the steps of the proof of Proposition 2.1 for some $s > 1$, we get a bound of the form

$$\|f\|_{\mathcal{B}_{\mathrm{RePU}_s}} \lesssim \left(\|\phi\|_{C^s(\mathbb{R})} + \|D^{s+1}\phi\|_{L^1(\mathbb{R})} \right) \int_\Omega (1 + \|w\|_{\ell^1} + |b|)^s d|\mu|(w,b) \tag{2.37}$$

for all $\mu \in \mathbb{G}_{\phi,f}$ for $f \in \mathcal{B}_\phi$. If we want an embedding, then we need to get rid of the exponent s in the integral on the right-hand side or we need to introduce the exponent in the norm of $\mathcal{B}_\phi$. In the following proposition we show the former by showing an embedding holds under the assumption that Ω is bounded. In Section 5 we briefly discuss different exponents in the Barron norm.

Proposition 2.2. *Let $s \in \mathbb{N}$ and Ω bounded. If $\phi \in C^s(\mathbb{R})$ such that $D^{s+1}\phi \in L^1(\mathbb{R})$, then $\mathcal{B}_\phi \hookrightarrow \mathcal{B}_{\mathrm{RePU}_s}$.*

Proof. Let $\mu \in \mathbb{G}_{\phi,f}$ for $f \in \mathcal{B}_\phi$, and recall that from the Taylor remainder theorem it follows that for given $(w,b) \in \Omega$

$$\phi(\langle x|w\rangle + b) = \sum_{k=0}^s \frac{\partial^k \phi(0)}{k!}(\langle x|w\rangle + b)^k + \int_0^{\langle x|w\rangle + b} \frac{D^{s+1}\phi(t)}{s!}(\langle x|w\rangle + b - t)^s dt \tag{2.38}$$

for all $x \in \mathcal{X}$.

For the series part, there exists a measure $\nu_{series} \in \mathcal{M}(\mathbb{R}^2)$ according to Lemma 2.2 such that

$$\int_{\mathbb{R}^2} \mathrm{RePU}_s(\omega z + \beta) d\nu_{series}(\omega, \beta) = \sum_{k=0}^s \frac{\partial \phi^k(0)}{k!} z^k \tag{2.39}$$

for all $z \in \mathbb{R}$ and

$$\int_{\mathbb{R}^2} (\|\omega\|_{\ell^1} + |\beta|)^s d|\nu_{series}|(\omega, \beta) \lesssim \|\phi\|_{C^s(\mathbb{R})}. \tag{2.40}$$

Observe that

$$\int_\Omega \sum_{k=0}^s \frac{\partial^k \phi(0)}{k!}(\langle x|w\rangle + b)^k d\mu(w,b) = \int_\Omega \int_{\mathbb{R}^2} \mathrm{RePU}_s(\omega(\langle x|w\rangle + b) + \beta) d\nu_{series}(\omega, \beta) d\mu(w,b)$$

$$= \int_\Omega \int_{\mathbb{R}^2} \mathrm{RePU}_s(\langle x|\omega w\rangle + \omega b + \beta) d\nu_{series}(\omega\beta) d\mu(w,b) \tag{2.41}$$

$$= \int_\Omega \mathrm{RePU}_s(\langle x|w\rangle + b) d\gamma_{series}(w,b)$$

where $\gamma_{series} := \Theta_\#(\nu_{series} \otimes \mu)$ is the push-forward along the map

$$\Theta : \Omega \times \mathbb{R}^2 \to \Omega, \quad ((w,b),(\omega,\beta)) \mapsto (\omega w, \omega b + \beta). \tag{2.42}$$

Furthermore,

$$\int_\Omega (\|w\|_{\ell^1} + |b|)^s d|\gamma_{series}|(w,b) \leq \int_\Omega \int_{\mathbb{R}^2} (\|\omega w\| + |\omega b + \beta|)^s d|\nu_{series}|(\omega, \beta) d|\mu|(w,b)$$

$$\leq \int_\Omega \int_{\mathbb{R}^2} (|\omega|\|w\| + |\omega||b| + |\beta|)^s d|\nu_{series}|(\omega, \beta) d|\mu|(w,b)$$

$$\leq \int_{\mathbb{R}^2} (|\omega| + |\beta|)^s d|\nu_{series}|(\omega, \beta) \int_\Omega (1 + \|w\|_{\ell^1} + |b|)^s d|\mu|(w,b) \tag{2.43}$$

$$\lesssim \|\phi\|_{C^s(\mathbb{R})} \int_\Omega (1 + \|w\|_{\ell^1} + |b|) d|\mu|(w,b).$$

For the remainder part, observe that

$$|\langle x|w\rangle + b| \leq \|w\| + |b| := \theta_{w,b} \tag{2.44}$$

for all $x \in \mathcal{X}$ and $(w,b) \in \Omega$. We can write that according to Lemma 2.3 as

$$\int_0^{\langle x|w\rangle+b} D^{s+1}\phi(u)(\langle x|w\rangle + b - u)^s du = \int_0^{\theta_{w,b}} D^{s+1}\phi(u)\,\mathrm{RePU}_s(\langle x|w\rangle + b - u) \tag{2.45}$$
$$+ (-1)^{s-1}D^{s+1}\phi(-u)\,\mathrm{RePU}_s(\langle x|-w\rangle - b - u)du.$$

After the change of coordinates $u = \theta_{w,b}t$, we get

$$\int_0^{\langle x|w\rangle+b} D^{s+1}\phi(u)(\langle x|w\rangle + b - u)^s du = \int_0^1 D^{s+1}\phi(\theta_{w,b}t)\,\mathrm{RePU}_s(\langle x|w\rangle + b - \theta_{w,b}t) \tag{2.46}$$
$$+ (-1)^{s-1}D^{s+1}\phi(-\theta_{w,b}t)\,\mathrm{RePU}_s(\langle x|-w\rangle - b - \theta_{w,b}t)\theta_{w,b}dt.$$

Observe that

$$\int_\Omega \int_0^{\langle x|w\rangle+b} D^{s+1}\phi(u)(\langle x|w\rangle + b - u)^s du d\mu(w,b) = \int_\Omega \mathrm{RePU}_s(\langle x|w\rangle + b)d\gamma_{rem}(w,b), \tag{2.47}$$

where the measure $\gamma_{rem} = \gamma_1 + \gamma_2$ is the sum of measures formed from the measures

$$\gamma_1(A) = \int_A \int_0^1 \theta_{w,b} D^{s+1}\phi(\theta_{w,b}u)d\Theta^1_\#(\mu \otimes \lambda)((w,b),u) \tag{2.48}$$
$$\gamma_2(A) = \int_A \int_0^1 \theta_{w,b} D^{s+1}\phi(-\theta_{w,b}u)d\Theta^2_\#(\mu \otimes \lambda)((w,b),u)$$

for Borel sets $A \subseteq \Omega$ and using the push-forward maps

$$\Theta^1 : \mathbb{R}^{d+1} \times [0,1] \to \mathbb{R}^{d+1}, \;\; ((w,b),u) \mapsto (w, b - \theta_{w,b}u) \tag{2.49}$$
$$\Theta^2 : \mathbb{R}^{d+1} \times [0,1] \to \mathbb{R}^{d+1}, \;\; ((w,b),u) \mapsto (-w, -b - \theta_{w,b}u),$$

where λ is the Lebesgue measure on $[0,1]$. Furthermore,

$$\int_\Omega (\|w\|_{\ell^1} + |b|)^s d|\gamma_{rem}|(w,b) \leq \int_\Omega \int_{-1}^1 \theta_{w,b}|D^{s+1}\phi(\theta_{w,b}t)|(\|w\|_{\ell^1} + |b - \theta_{w,b}t|)^s dt d|\mu|(w,b)$$
$$\leq \int_\Omega \int_{-1}^1 \theta_{w,b}|D^{s+1}\phi(\theta_{w,b}t)|(\|w\|_{\ell^1} + |b| + \theta_{w,b}|t|)^s dt d|\mu|(w,b)$$
$$= \int_\Omega \int_{-1}^1 \theta_{w,b}|D^{s+1}\phi(\theta_{w,b}t)|(1 + |t|)^s(\|w\|_{\ell^1} + |b|)^s dt|\mu|(w,b) \tag{2.50}$$
$$\lesssim \int_\Omega \int_{-\theta_{w,b}}^{\theta_{w,b}} |D^{s+1}\phi(u)|(\|w\|_{\ell^1} + |b|)^s du|\mu|(w,b)$$
$$\leq \|D^{s+1}\phi\|_{L^1(\mathbb{R})} \int_\Omega (1 + \|w\|_{\ell^1} + |b|)^s d|\mu|(w,b)$$
$$\lesssim \|D^{s+1}\phi\|_{L^1(\mathbb{R})} \int_\Omega (1 + \|w\|_{\ell^1} + |b|)d|\mu|(w,b).$$

Hence, we get by combining the remainder part with the series that $\gamma_{rem} + \gamma_{series} \in \mathbb{G}_{\mathrm{RePU}_s,f}$ with bound

$$\|f\|_{\mathcal{B}_{\mathrm{RePU}_s}} \lesssim (\|\phi\|_{C^s(\mathbb{R})} + \|D^{s+1}\phi\|_{L^1(\mathbb{R})})\|f\|_{\mathcal{B}_\phi} \tag{2.51}$$

where we took the infimum over $\mu \in \mathbb{G}_{\phi,f}$. $\qquad Q.E.D.$

2.3 Hierarchy in the RePU

In Proposition 2.2 we have shown an embedding into $\mathcal{B}_{\mathrm{RePU}_s}$ using a smoothness criterion. For the continuous functions it is well-known that they have a hierarchical structure, i.e. $C^s \hookrightarrow C^t$ for all $s, t \in \mathbb{N}$ such that $t \leq s$. The following proposition shows that $\mathcal{B}_{\mathrm{RePU}_s}$ has a similar hierarchy. It is a generalization of 1) from Lemma 7.1 of [Caragea et al., 2020].

Proposition 2.3. *For $s, t \in \mathbb{N}$ with $t \leq s$ we have $\mathcal{B}_{\mathrm{RePU}_s} \hookrightarrow \mathcal{B}_{\mathrm{RePU}_t}$.*

Proof. Let $c > 0$. A relation between RePU_t and RePU_{t+1} is given by

$$\mathrm{RePU}_{t+1}(y) = (t+1) \int_0^c \mathrm{RePU}_t(y-u)du \qquad (2.52)$$

for all $y \in \mathbb{R}$ with $\|y\| \leq c$. We will use this relation to prove that

$$\mathcal{B}_{\mathrm{RePU}_{t+1}} \hookrightarrow \mathcal{B}_{\mathrm{RePU}_t}. \qquad (2.53)$$

The proposition follows from (2.53).

Let $\mu \in \mathbb{G}_{\mathrm{RePU}_{t+1}, f}$ for $f \in \mathcal{B}_{\mathrm{RePU}_{t+1}}$. Observe that

$$
\begin{aligned}
f(x) &= \int_{\mathbb{R}^{d+1}} \mathrm{RePU}_{t+1}(\langle x|w \rangle + b) d\mu(w, b) \\
&= \int_{\mathbb{R}^{d+1}} (t+1) \int_0^{\theta_{w,b}} \mathrm{RePU}_t(\langle x|w \rangle + b - u) du \, d\mu(w, b) \qquad (2.52) \\
&= \int_{\mathbb{R}^{d+1}} \int_0^1 (t+1)\theta_{w,b} \mathrm{RePU}_t(\langle x|w \rangle + b - \theta_{w,b}v) dv \, d\mu(w, b) \quad u = \theta_{w,b}v \\
&= \int_{\mathbb{R}^{d+1}} \mathrm{RePU}_t(\langle x|w \rangle + b) d\nu(w, b),
\end{aligned} \qquad (2.54)
$$

where $\theta_{w,b} := \|w\|_{\ell^1} + |b|$, λ is the Lebesgue measure on $[0, 1]$ and the measure

$$\nu(A) = (t+1) \int_A \int_0^1 \theta_{w,b} d\Theta_{\#}(\mu \otimes \lambda)((w, b), v) \qquad (2.55)$$

for the Borel sets $A \subseteq \Omega$ is the push forward of $\mu \otimes \lambda$ along the map

$$\Theta : \mathbb{R}^{d+1} \times [0, 1] \to \mathbb{R}^{d+1}, \ ((w, b), v) \mapsto (w, b - \theta_{w,b}v). \qquad (2.56)$$

Hence, $\nu \in \mathbb{G}_{\mathrm{RePU}_t, f}$. Furthermore,

$$
\begin{aligned}
\|f\|_{\mathcal{B}_{\mathrm{RePU}_t}} &\leq \int_{\mathbb{R}^{d+1}} (\|w\|_{\ell^1} + |b|)^t d|\nu|(w, b) \\
&\leq \int_{\mathbb{R}^{d+1}} \int_0^1 (t+1)\theta_{w,b}(\|w\|_{\ell^1} + |b - \theta_{w,b}v|)^t dv \, d|\mu|(w, b) \\
&\leq \int_{\mathbb{R}^{d+1}} \int_0^1 (t+1)\theta_{w,b}(\|w\|_{\ell^1} + |b| + \theta_{w,b}|v|)^t dv \, d|\mu|(w, b) \\
&= \int_{\mathbb{R}^{d+1}} \int_0^1 (t+1)(\|w\|_{\ell^1} + |b|)^{t+1}(1+v)^t dv \, d|\mu|(w, b) \\
&= (t+1)\frac{2^{t+1} - 1}{t+1} \int_{\mathbb{R}^{d+1}} (\|w\|_{\ell^1} + |b|)^{t+1} d|\mu|(w, b) \\
&= (2^{t+1} - 1) \int_{\mathbb{R}^{d+1}} (\|w\|_{\ell^1} + |b|)^{t+1} d|\mu|(w, b).
\end{aligned} \qquad (2.57)
$$

Taking the infimum over $\mu \in \mathbb{G}_{\mathrm{RePU}_{t+1}, f}$ gives

$$\|f\|_{\mathcal{B}_{\mathrm{RePU}_t}} \leq (2^{t+1} - 1)\|f\|_{\mathcal{B}_{\mathrm{RePU}_{t+1}}}. \qquad (2.58)$$

Q.E.D.

3   Embeddings between Barron spaces for Lipschitz activation functions

In the previous sections, we dealt with the relations between two Barron spaces when one of the Barron spaces had RePU_s as the activation function. In this section, we take a broader perspective and look

at the relations between Barron spaces with Lipschitz activation functions. In particular, we provide embeddings when the activation functions can be written as a linear combination of scaled and shifted versions of another activation function and as a convolution with another activation function in Section 3.1 and when one of the two activation functions is the derivative of the other in Section 3.2.

3.1 Activation functions related by linear combinations and convolutions

Let ϕ be a Lipschitz activation function. Since the values for the weights and biases are not restricted a priori, we expect that a Barron space with activation function

$$\psi(x) = c_0\phi(c_1 x + c_2) \tag{3.1}$$

for some $c_0, c_1 \neq 0$ and $c_2 \in \mathbb{R}$ is similar to that with ϕ. At the same time, we expect that a Barron space with

$$\psi(x) = \phi(x) - \phi(x - c_3) \tag{3.2}$$

for $c_3 \neq 0$ embeds in that with ϕ. Both of these and more are covered by the following proposition.

Proposition 3.1. *If ψ and ϕ are Lipschitz activation functions such that*

$$\phi(x) = \int_{\mathbb{R}^2} \psi(\omega x + \beta) d\gamma(\omega, \beta) \tag{3.3}$$

for some measure $\gamma \in \mathcal{M}(\mathbb{R}^2)$ satisfying

$$\int_{\mathbb{R}^2} (1 + |\omega| + |\beta|) d|\gamma|(\omega, \beta) < \infty, \tag{3.4}$$

then $\mathcal{B}_\phi \hookrightarrow \mathcal{B}_\psi$.

Proof. Let $\mu \in \mathbb{G}_{\phi,f}$ for some $f \in \mathcal{B}_\phi$. This means that

$$
\begin{aligned}
f(x) &= \int_{\mathbb{R}^{d+1}} \phi(\langle x|w \rangle + b) d\mu(w, b) \\
&= \int_{\mathbb{R}^{d+1}} \int_{\mathbb{R}^2} \psi(\omega(\langle x|w \rangle + b) + \beta) d\gamma(\omega, \beta) d\mu(w, b) \\
&= \int_{\mathbb{R}^{d+1} \times \mathbb{R}^2} \psi(\langle x|\omega w \rangle + \omega b + \beta) d(\gamma \otimes \mu)((\omega, \beta), (w, b)) \\
&= \int_{\mathbb{R}^{d+1}} \psi(\langle x|w \rangle + b) d\nu(w, b)
\end{aligned} \tag{3.5}
$$

where the measure $\nu := \Theta_{\#}(\gamma \otimes \mu)$ is the push forward along the map

$$\Theta : \Omega \times \mathbb{R}^2 \to \Omega, \ (\omega, \beta), (w, b)) \to (\omega w, \omega b + \beta). \tag{3.6}$$

Hence, $\nu \in \mathbb{G}_{\psi,f}$. Furthermore,

$$
\begin{aligned}
\|f\|_{\mathcal{B}_\psi} &\leq \int_{\mathbb{R}^{d+1}} (1 + \|w\|_{\ell^1} + |b|) d|\nu|(w, b) \\
&\leq \int_{\mathbb{R}^{d+1}} \int_{\mathbb{R}^2} (1 + \|\omega w\| + |\omega b + \beta|) d|\gamma|(\omega, \beta) d|\mu|(w, b) \\
&\leq \int_{\mathbb{R}^{d+1}} \int_{\mathbb{R}^2} (1 + |\omega| \|w\|_{\ell^1} + |\omega||b| + |\beta|) d|\gamma|(\omega, \beta) d|\mu|(w, b) \\
&\leq \int_{\mathbb{R}^{d+1}} \int_{\mathbb{R}^2} (1 + |\omega| + |\beta|)(1 + \|w\|_{\ell^1} + |b|) d|\gamma|(\omega, \beta) d|\mu|(w, b) \\
&= \int_{\mathbb{R}^2} (1 + |\omega| + |\beta|) d|\gamma|(\omega, \beta) \int_{\mathbb{R}^{d+1}} (1 + \|w\|_{\ell^1} + |b|) d|\mu|(w, b).
\end{aligned} \tag{3.7}
$$

Taking the infimum over $\mu \in \mathbb{G}_{\psi, f}$ gives

$$\|f\|_{\mathcal{B}_\psi} \leq \int_{\mathbb{R}^2} (1 + |\omega| + |\beta|) d|\gamma|(\omega, \beta) \|f\|_{\mathcal{B}_\phi}. \tag{3.8}$$

$$Q.E.D.$$

Informally, Proposition 3.1 says that if ϕ is an element of the Barron space with ψ as the activation function for $d = 1$, then the embedding of $\mathcal{B}_\phi$ into $\mathcal{B}_\psi$ holds. This not only covers the aforementioned cases in (3.1) and (3.2) but also when ϕ is a convolution of ψ with some kernel η or when ϕ can be written as a series expansion in ψ.

Corollary 3.1. *If $\eta : \mathbb{R} \to \mathbb{R}$ satisfies*

$$\int_{\mathbb{R}} |\eta(z)|(1 + |z|)dz \leq C \tag{3.9}$$

and ϕ is a Lipschitz activation function, then $\mathcal{B}_{\phi * \eta} \hookrightarrow \mathcal{B}_\phi$ with

$$\|f\|_{\mathcal{B}_\phi} \leq C\|f\|_{\mathcal{B}_{\phi * \eta}} \tag{3.10}$$

for all $f \in \mathcal{B}_{\phi * \eta}$.

Corollary 3.2. *Let ϕ and ψ be two Lipschitz activation functions linked by*

$$\phi(x) = \sum_{k=1}^{\infty} g(k)\psi(h(k)x) \tag{3.11}$$

with

$$\sum_{k=1}^{\infty} |g(k)|(1 + |h(k)|) \leq C, \tag{3.12}$$

then $\mathcal{B}_\phi \hookrightarrow \mathcal{B}_\psi$ with

$$\|f\|_{\mathcal{B}_\psi} \leq C\|f\|_{\mathcal{B}_\phi} \tag{3.13}$$

for all $f \in \mathcal{B}_\phi$.

Note that Corollary 3.1 is particularly convenient if one knows the Fourier transforms of the relevant activation functions ϕ and ψ. In that case, it is sufficient to check whether the kernel η defined using its Fourier transform

$$\hat{\eta} := \frac{\hat{\phi}}{\hat{\psi}} \tag{3.14}$$

satisfies the growth condition. This allows one for example to show that

$$\mathcal{B}_{\tanh} \hookrightarrow \mathcal{B}_{\arctan}. \tag{3.15}$$

3.2 Activation functions related by a derivative

Something that Proposition 3.1 does not cover, is when one activation function is the derivative of another. An example of this is

$$SoftPlus(x) = \log(1 + e^x) \tag{3.16}$$

and

$$logi(x) = \frac{1}{1 + e^{-x}} \tag{3.17}$$

related by

$$\partial SoftPlus(x) = logi(x). \tag{3.18}$$

In this case, only an inclusion has been found and not an embedding.

Proposition 3.2. *If ζ is a continuously differentiable activation function with $Lip(\zeta) < \infty$, then $\mathcal{B}_\zeta \subseteq \mathcal{B}_{\partial\zeta}$.*

Proof. Let $\mu \in \mathbb{G}_{\partial\zeta,f}$ for $f \in \mathcal{B}_{\partial\zeta}$. Consider the sequence of measures $\{v^h\}_{h>0}$ given by

$$\nu^h = \frac{1}{h}\left(\Theta^h_\#\mu - \mu\right) \tag{3.19}$$

along the map

$$\Theta^h : \mathbb{R}^{d+1} \to \mathbb{R}^{d+1}, \ (w,b) \mapsto (w, b+h). \tag{3.20}$$

Observe that for

$$f_h(x) = \int_{\mathbb{R}^{d+1}} \zeta(\langle x|w\rangle + b)d\nu^h(w,b) \tag{3.21}$$

we have

$$
\begin{aligned}
\|f_h\|_{\mathcal{B}_\zeta} &\leq \frac{1}{h}\int_{\mathbb{R}^{d+1}}(1 + \|w\|_{\ell^1} + |b|)d|\Theta^h_\#\mu - \mu|(w,b) \\
&\leq \frac{1}{h}\int_{\mathbb{R}^{d+1}}(1 + \|w\|_{\ell^1} + |b|)d\left(|\Theta^h_\#\mu| + |\mu|\right)(w,b) \\
&= \frac{1}{h}\int_{\mathbb{R}^{d+1}} 2(1 + \|w\|_{\ell^1}) + |b+h| + |b|d|\mu|(w,b) \\
&\leq \frac{1}{h}\int_{\mathbb{R}^{d+1}} 2(1 + \|w\|_{\ell^1} + |b|) + |h|d|\mu|(w,b) \\
&\leq \frac{2+h}{h}\int_{\mathbb{R}^{d+1}} 1 + \|w\|_{\ell^1} + |b|d|\mu|(w,b) \\
&< \infty
\end{aligned}
\tag{3.22}
$$

for all $h > 0$. Hence, $f_h \in \mathcal{B}_\zeta$. The sequence $\{f_h\}_{h>0}$ satisfies

$$
\begin{aligned}
\lim_{h\to 0} f^h(x) &= \lim_{h\to 0}\int_{\mathbb{R}^{d+1}} \zeta(\langle x|w\rangle + b)d\nu^h(w,b) \\
&= \lim_{h\to 0}\int_{\mathbb{R}^{d+1}} \frac{\zeta(\langle x|w\rangle + b + h) - \zeta(\langle x|w\rangle + b)}{h}d\mu(w,b) \\
&= \int_{\mathbb{R}^{d+1}} \lim_{h\to 0} \frac{\zeta(\langle x|w\rangle + b + h) - \zeta(\langle x|w\rangle + b)}{h}d\mu(w,b) \quad \text{Dominated conv. th.} \\
&= \int_{\mathbb{R}^{d+1}} \partial\zeta(\langle x|w\rangle + b)d\mu(w,b) \\
&= f(x),
\end{aligned}
\tag{3.23}
$$

where we are allowed to use the dominated convergence theorem since

$$\left|\frac{\zeta(\langle x|w\rangle + b + h) - \zeta(\langle x|w\rangle + b)}{h}\right| \leq Lip(\zeta) < \infty. \tag{3.24}$$

Since the Barron space $\mathcal{B}_\zeta$ is complete and $f_h \to f$, we have that $f \in \mathcal{B}_\zeta$.   *Q.E.D.*

4   Embeddings for spectral Barron spaces

We have shown that embeddings between different Barron spaces can be proven by constructing suitable the push-forwards. This strategy can also be used to show embeddings between a Barron space and a non-Barron space. We will demonstrate this in this section by showing the embedding of the spectral Barron spaces into the Barron spaces with a $RePU_s$ as activation function. This embedding is a generalization of 3) from Lemma 7.1 of [Caragea et al., 2020].

We recall that the spectral Barron spaces are given by

$$\mathbb{G}_{s,f}^{\mathscr{F}} = \left\{ f \in L^1([-1,1]^d) \;\middle|\; \exists f_e \in L^1(\mathbb{R}^d): \; f_e|_{\mathcal{X}} = f \right\}$$

$$\|f\|_{\mathcal{B}_{\mathscr{F},s}} = \inf_{f_e \in \mathbb{G}_{s,f}^{\mathscr{F}}} \int_{\mathbb{R}^d} (1 + \|\xi\|_{\ell^1})^s \left| \hat{f}_e(\xi) \right| d\xi \tag{4.1}$$

$$\mathcal{B}_{\mathscr{F},s} = \left\{ f: L^1([-1,1]^d) \;\middle|\; \|f\|_{\mathcal{B}_{\mathscr{F},s}} < \infty \right\}$$

for $s \in \mathbb{N}$, and have been called Spectral spaces and Auxiliary spaces as well. From Lemma 2.7 of [Voigtlaender, 2022] it follows that for each $s \in \mathbb{N}$ all the functions $f \in \mathcal{B}_{\mathscr{F},s+1}$ satisfy the conditions for the multivariate Taylor remainder theorem, i.e.

$$f(x) = \sum_{|\alpha| \leq s} \frac{\partial^\alpha f(0)}{\alpha!} x^\alpha + \sum_{|\alpha|=s+1} \frac{s+1}{\alpha!} x^\alpha \int_0^1 (1-t)^s D^\alpha f(tx) dt, \tag{4.2}$$

holds, where we have used the multi-index notation for α. Similar to the univariate case, we can use to Lemma 2.2 to construct a suitable push-forward map for the series part. However, unlike the univariate case, there is no analogue to Lemma 2.3 to help us construct a push-forward map for the remainder part. Fortunately, we can construct one by using the spectral nature of $f \in \mathcal{B}_{\mathscr{F},s+1}$.

Proposition 4.1. *For all $s \in \mathbb{N}$ it holds that $\mathcal{B}_{\mathscr{F},s+1} \hookrightarrow \mathcal{B}_{\mathrm{RePU}_s}$.*

Proof. Let $f_e \in \mathbb{G}_{s,f}^{\mathscr{F}}$ for $f \in \mathcal{B}_{\mathscr{F},s+1}$. Recall that

$$f(x) = \int_{\mathbb{R}^d} \mathrm{e}^{i\langle x|\xi\rangle} \hat{f}_e(\xi) d\xi \tag{4.3}$$

for all $x \in \mathcal{X}$. The integral form of the Taylor remainder theorem for the exponential map $z \mapsto \mathrm{e}^{iz}$ around the origin up to order s is given by

$$\mathrm{e}^{iz} = \sum_{k=0}^s \frac{i^k}{k!} z^k + \int_0^z \frac{i^{s+1} \mathrm{e}^{it}}{s!} (z-t)^s dt. \tag{4.4}$$

Substituting (4.4) into the right-hand side of (4.3) gives

$$\int_{\mathbb{R}^d} \mathrm{e}^{i\langle x|\xi\rangle} \hat{f}_e(\xi) d\xi = \int_{\mathbb{R}^d} \sum_{k=0}^s \frac{i^k}{k!} \langle x|\xi\rangle^k \hat{f}_e(\xi) d\xi + \int_{\mathbb{R}^d} \int_0^{\langle x|\xi\rangle} \frac{i^{s+1} \mathrm{e}^{it}}{s!} (\langle x|\xi\rangle - t)^s dt \hat{f}_e(\xi) d\xi. \tag{4.5}$$

For the series part, we observe that by the Fourier derivation identity

$$\int_{\mathbb{R}^d} \sum_{k=0}^s \frac{i^k}{k!} \langle x|\xi\rangle^k \hat{f}_e(\xi) d\xi = \sum_{|\alpha| \leq s} \frac{\partial^\alpha f(0)}{\alpha!} x^\alpha. \tag{4.6}$$

For the remainder part, we observe that $|\langle x|\xi\rangle| \leq \|\xi\|_{\ell^1}$, thus by Lemma 2.3

$$\int_0^{\langle x|\xi\rangle} \frac{i^{s+1} \mathrm{e}^{it}}{s!} (x-t)^s dt = \frac{i^{s+1}}{s!} \int_0^{\|\xi\|_{\ell^1}} \left(\mathrm{e}^{it} \, \mathrm{RePU}_s(\langle x|\xi\rangle - t) + (-1)^{s-1} \mathrm{e}^{-it} \, \mathrm{RePU}_s(-\langle x|\xi\rangle - t) \right) dt. \tag{4.7}$$

After doing the change of coordinates $u = \|\xi\|_{\ell^1} t$ and substituting the resultant expression together with (4.6) into the right-hand side of (4.5), we get

$$f(x) = \sum_{|\alpha| \leq s} \frac{\partial^\alpha f(0)}{\alpha!} x^\alpha + \int_{\mathbb{R}^d} \int_0^1 \frac{i^{s+1} \|\xi\|_{\ell^1}^{s+1} \hat{f}_e(\xi)}{s!} \left(\mathrm{e}^{i\|\xi\|_{\ell^1} u} \, \mathrm{RePU}_s \left(\left\langle x \middle| \frac{\xi}{\|\xi\|_{\ell^1}} \right\rangle - u \right) \right.$$

$$\left. + \mathrm{e}^{-i\|\xi\|_{\ell^1} u} \, \mathrm{RePU}_s \left(\left\langle x \middle| \frac{-\xi}{\|\xi\|_{\ell^1}} \right\rangle - u \right) \right) du d\xi. \tag{4.8}$$

We can remove the second RePU_s term by observing that

$$\int_{\mathbb{R}^d} \int_0^1 \frac{i^{s+1}\|\xi\|_{\ell^1}^{s+1}\hat{f}_e(\xi)}{s!}e^{-i\|\xi\|_{\ell^1}u}\,\text{RePU}_s\left(\left\langle x\middle|\frac{-\xi}{\|\xi\|_{\ell^1}}\right\rangle - u\right)dud\xi$$
$$= -\int_{\mathbb{R}^d}\int_0^1 \frac{i^{s+1}\|\xi\|_{\ell^1}^{s+1}\hat{f}_e(-\xi)}{s!}e^{-i\|\xi\|_{\ell^1}u}\,\text{RePU}_s\left(\left\langle x\middle|\frac{\xi}{\|\xi\|_{\ell^1}}\right\rangle - u\right)dud\xi, \tag{4.9}$$

where we used the coordinate map $\xi \mapsto -\xi$. Substituting (4.9) into (4.8) gives

$$f(x) = \sum_{|\alpha|\le s}\frac{\partial^\alpha f(0)}{\alpha!}x^\alpha + \int_{\mathbb{R}^d}\int_0^1 \frac{i^{s+1}\|\xi\|_{\ell^1}^{s+1}}{s!}\left(\hat{f}_e(\xi)e^{i\|\xi\|_{\ell^1}u} - \hat{f}_e(-\xi)e^{-i\|\xi\|_{\ell^1}u}\right)\text{RePU}_s\left(\left\langle x\middle|\frac{\xi}{\|\xi\|_{\ell^1}}\right\rangle - u\right)dud\xi. \tag{4.10}$$

From Lemma 2.2 it follows that there exists a measure $\mu_{series} \in \mathcal{M}(\Omega)$ such that

$$\sum_{|\alpha|\le s}\frac{\partial^\alpha f(0)}{\alpha!}x^\alpha = \int_\Omega \text{RePU}_s(\langle x|w\rangle + b)d\mu_{series}(w,b). \tag{4.11}$$

Simultaneously, we observe that

$$f(x) = \int_\Omega \text{RePU}_s(\langle x|w\rangle + b)d\mu(w,b), \tag{4.12}$$

where the measure $\mu := \mu_{series} + \mu_{rem}$ is the sum of the measure μ_{series} and the measure μ_{rem} given by

$$d\mu_{rem}(\xi,u) = \int_0^1 Re\left(\frac{i^{s+1}\|\xi\|_{\ell^1}^{s+1}}{s!}\left(\hat{f}_e(\xi)e^{i\|\xi\|_{\ell^1}u} - \hat{f}_e(-\xi)e^{-i\|\xi\|_{\ell^1}u}\right)\right)d\Theta_\#(\lambda_{\mathbb{R}^d}\otimes\lambda_{[0,1]})(\xi,u) \tag{4.13}$$

defined using the push-forward map

$$\Theta:\mathbb{R}^d \times [0,1] \to \S^d \times [0,1], \ (\xi,u) \mapsto (\frac{\xi}{\|\xi\|_{\ell^1}}, u) \tag{4.14}$$

with $\lambda_{\mathbb{R}^d}$ and $\lambda_{[0,1]}$ the Lebesgue measures on $\mathbb{R}^d$ and $[0,1]$ respectively. Hence, $\mu \in \mathbb{G}_{\text{RePU}_s,f}$. Furthermore,

$$\|f\|_{\mathcal{B}_{\text{RePU}_s}} \le \int_\Omega (\|w\|_{\ell^1} + |b|)^s d|\mu|(w,b)$$
$$\le \int_\Omega (\|w\|_{\ell^1} + |b|)^s d|\mu_{series}|(w,b) + \int_\Omega (\|w\|_{\ell^1} + |b|)^s d|\mu_{rem}|(w,b)$$
$$\lesssim \|f_e\|_{C_0^s(\mathbb{R}^d)} + \|\mu_{rem}\|_{\mathcal{M}(\Omega)}$$
$$\lesssim \left\|\hat{f}_e\right\|_{L^1(\mathbb{R}^d,(1+\|\cdot\|)^{s+1})}, \tag{4.15}$$

where we used Lemma 2.7 of [Voigtlaender, 2022] to bound $\|f_e\|_{C_0^s(\mathbb{R}^d)}$. Taking the infimum over $f_e \in \mathbb{G}_{s,f}^{\mathscr{F}}$ gives

$$\|f\|_{\mathcal{B}_{\text{RePU}_s}} \lesssim \|f\|_{\mathcal{B}_{\mathscr{F},s+1}}. \tag{4.16}$$

$$Q.E.D.$$

5   Discussion and conclusion

In this paper, we have studied the effect of changing the activation function on the Barron spaces. This has been done by determining embeddings between two Barron spaces with different activation functions.

We have shown that the Barron spaces with RePU_s have a hierarchical structure, i.e. if $t \le s$ for $t, s \in \mathbb{N}$, then the Barron space with RePU_s embeds into that with RePU_t. This structure is similar to well-known Sobolev spaces H^s and the continuous function spaces C^s. In [E and Wojtowytsch, 2022b], four PDEs with explicit formulas for their solutions are studied. These formulas can be derived using the Green's

function associated with the PDE. They discuss several challenges when using Barron functions for the initial conditions and/or boundary conditions. When using Sobolev spaces, many of these challenges are overcome by assuming higher regularity. Some remaining challenges can potentially be solved by assuming higher s for the Barron spaces with RePU_s.

The embeddings, for which we assume that neither is a RePU_s, cover many of the changes that are made to existing activation function in order to find new ones to use. Examples of such changes are scaling and shifting (compare *logi* with tanh), taking a linear combination (compare leaky ReLU with ReLU) and taking a derivative (compare *SoftPlus* with *logi*).

Although our results cover many activation functions, we have only provided affirmative statements, i.e. we provided statements that show a suitable push-forward map Θ exists and we provided no statements that show no such map can exist. Consider as an example the sawtooth wave function $\text{SawTooth}_{A,p}$ with amplitude A and period p as activation function. This function has been used to show the relevance of depth in neural networks with ReLU as the activation function [Telgarsky, 2015]. It has a series representation in terms of sin given by

$$\text{SawTooth}_{A,p}(x) = A\left(\frac{1}{2} - \frac{1}{\pi}\sum_{k=0}^{\infty}(-1)^k\frac{\sin(2\pi kpx)}{k}\right). \tag{5.1}$$

For every $\mu \in \mathbb{G}_{\text{SawTooth}_{A,p},f}$ for $f \in \mathcal{B}_{\text{SawTooth}_{A,p}}$ we can find a measure $\nu \in \mathcal{M}(\mathbb{R}^{d+1})$ such that

$$\int_{\mathbb{R}^{d+1}} \text{SawTooth}_{A,p}(\langle x|w\rangle + b)d\mu(w,b) = \int_{\mathbb{R}^{d+1}} \sin(\langle x|w\rangle + b)d\nu(w,b). \tag{5.2}$$

The right-hand side of (5.2) is a well-defined integral, but

$$\int_{\mathbb{R}^{d+1}}(1 + \|w\|_{\ell^1} + |b|)d|\nu|(w,b) \tag{5.3}$$

does not converge. Our results do not rule out the existence of an embedding of $\mathcal{B}_{\text{SawTooth}_{A,p}}$ into $\mathcal{B}_{\sin}$, yet they provide support that such an embedding may not exist at all.

Our results are also limited to RePU_s or Lipschitz activation functions. This restriction makes sure that, given an activation function σ, a function $f \in \mathcal{B}_\sigma$ of the form (1.1) is well-defined for all $\mu \in \mathbb{G}_\sigma$. An activation function like

$$\sigma(x) = |x|^2 \tag{5.4}$$

is not covered by this [Sarao Mannelli et al., 2020]. This activation function is asymptotically quadratic and is thus not Lipschitz. To cover activation functions like this the Barron spaces can be adapted by redefining the Barron norm for continuous non-homogeneous functions as

$$\|f\|_{\mathcal{B}_\sigma} = \inf_{\mu \in \mathbb{G}_{\sigma,f}}\int_\Omega (1 + \|w\|_{\ell^1} + |b|)^p d|\mu|(w,b) \tag{5.5}$$

with

$$p = \arg\min\left\{q \in \mathbb{N} \;\middle|\; \forall x \in \mathcal{X}, (w,b) \in \Omega : \; \frac{|\sigma(\langle x|w\rangle + b)|}{(1 + \|w\|_{\ell^1} + |b|)^q} < \infty\right\}. \tag{5.6}$$

When σ is Lipschitz continuous, $p = 1$. Hence, this recovers the Barron norm in that case. Note that results like Proposition 2.1 are still preserved, since for all $p \in \mathbb{N}$ we have

$$\inf_{\mu \in \mathbb{G}_{\sigma,f}}\int_\Omega (1 + \|w\|_{\ell^1} + |b|)^p d|\mu|(w,b) \geq \inf_{\mu \in \mathbb{G}_{\sigma,f}}\int_\Omega (1 + \|w\|_{\ell^1} + |b|)d|\mu|(w,b). \tag{5.7}$$

References

Bartolucci, F., De Vito, E., Rosasco, L., & Vigogna, S. (2023). Understanding neural networks with reproducing kernel Banach spaces. *Applied and Computational Harmonic Analysis*, *62*, 194–236. https://doi.org/10.1016/j.acha.2022.08.006

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

Chen, Q., Hao, W., & He, J. (2022). Power Series Expansion Neural Network [arXiv:2102.13221 [cs, math]]. *Journal of Computational Science*, *59*, 101552. https://doi.org/10.1016/j.jocs.2021.101552

DeVore, R., Hanin, B., & Petrova, G. (2021). Neural network approximation. *Acta Numerica*, *30*, 327–444. https://doi.org/10.1017/S0962492921000052

E, W., & Wojtowytsch, S. (2021). Kolmogorov width decay and poor approximators in machine learning: Shallow neural networks, random feature models and neural tangent kernels. *Research in the Mathematical Sciences*, *8*(1), 5. https://doi.org/10.1007/s40687-020-00233-4

E., W., & Wojtowytsch, S. (2022a). Representation formulas and pointwise properties for Barron functions. *Calculus of Variations and Partial Differential Equations*, *61*(2), 46. https://doi.org/10.1007/s00526-021-02156-6

E, W., & Wojtowytsch, S. (2022b). Some observations on high-dimensional partial differential equations with Barron data [ISSN: 2640-3498]. *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, 253–269. Retrieved May 19, 2023, from https://proceedings.mlr.press/v145/e22a.html

Gantmacher, F. R. (2009). *The theory of matrices. Vol. 2* (Reprinted, Vol. 2). American Mathematical Soc.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep Sparse Rectifier Neural Networks [ISSN: 1938-7228]. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 315–323. Retrieved December 13, 2022, from https://proceedings.mlr.press/v15/glorot11a.html

Hendrycks, D., & Gimpel, K. (2020). Gaussian Error Linear Units (GELUs) [arXiv:1606.08415 [cs] version: 4]. https://doi.org/10.48550/arXiv.1606.08415

Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L.-C., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H., & Le, Q. (2019). Searching for MobileNetV3. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1314–1324. https://doi.org/10.1109/ICCV.2019.00140

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications [arXiv:1704.04861 [cs] version: 1]. https://doi.org/10.48550/arXiv.1704.04861

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

Maas, A. L. (2013). Rectifier Nonlinearities Improve Neural Network Acoustic Models. Retrieved December 13, 2022, from https://www.semanticscholar.org/paper/Rectifier-Nonlinearities-Improve-Neural-Network-Maas/367f2c63a6f6a10b3b64b8729d601e69337ee3cc

Misra, D. (2020). Mish: A Self Regularized Non-Monotonic Activation Function [arXiv:1908.08681 [cs, stat] version: 3]. https://doi.org/10.48550/arXiv.1908.08681

Noel, M. M., L, A., Trivedi, A., & Dutta, P. (2021). *Growing Cosine Unit: A Novel Oscillatory Activation Function That Can Speedup Training and Reduce Parameters in Convolutional Neural Networks* (tech. rep. No. arXiv:2108.12943) (arXiv:2108.12943 [cs] type: article). arXiv. https://doi.org/10.48550/arXiv.2108.12943

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

Ramachandran, P., Zoph, B., & Le, Q. V. (2023). Searching for Activation Functions. Retrieved May 19, 2023, from https://openreview.net/forum?id=Hkuq2EkPf

Sarao Mannelli, S., Vanden-Eijnden, E., & Zdeborová, L. (2020). Optimization and Generalization of Shallow Neural Networks with Quadratic Activation Functions. *Advances in Neural Information Processing Systems*, *33*, 13445–13455. Retrieved January 24, 2023, from https://proceedings.neurips.cc/paper/2020/hash/9b8b50fb590c590ffbf1295ce92258dc-Abstract.html

Siegel, J. W., & Xu, J. (2020). Approximation rates for neural networks with general activation functions. *Neural Networks*, *128*, 313–321. https://doi.org/10.1016/j.neunet.2020.05.019

Siegel, J. W., & Xu, J. (2021). High-Order Approximation Rates for Neural Networks with ReLU^k Activation Functions [arXiv: 2012.07205]. *arXiv:2012.07205 [cs, math]*. Retrieved August 18, 2021, from http://arxiv.org/abs/2012.07205

Spek, L., Heeringa, T. J., Schwenninger, F., & Brune, C. (2023). Duality for Neural Networks through Reproducing Kernel Banach Spaces [arXiv:2211.05020 [cs, math]]. https://doi.org/10.48550/arXiv.2211.05020

Telgarsky, M. (2015). Representation Benefits of Deep Feedforward Networks [arXiv:1509.08101 [cs]]. https://doi.org/10.48550/arXiv.1509.08101

Voigtlaender, F. (2022). L^p sampling numbers for the Fourier-analytic Barron space [arXiv:2208.07605 [cs, math, stat]]. https://doi.org/10.48550/arXiv.2208.07605

Appendices

Proposition 2.1 does not cover piece-wise smooth activation functions. This is an alteration to some of these.

Proposition .1. *If ϕ is continuous and smooth everywhere except at the origin and*

$$\tilde{\gamma}(\phi) = \left(|\phi(0)| + \left|\partial^1\phi_+(0)\right| + \left|\partial^1\phi_-(0)\right| + 2\int_0^\infty \left|\partial^2\phi_+(z)\right|dz + 2\int_{-\infty}^0 \left|\partial^2\phi_-(z)\right|dz \right) < \infty, \qquad (.8)$$

then $\mathcal{B}_\phi \hookrightarrow \mathcal{B}_{\mathrm{ReLU}}$ with

$$\|f\|_{\mathcal{B}_{\mathrm{ReLU}}} \le \tilde{\gamma}(\phi)\|f\|_{\mathcal{B}_\phi} \qquad (.9)$$

for all $f \in \mathcal{B}_\phi$.

Proof. From the assumptions on ϕ, it follows that it can be written as

$$\phi(x) = \begin{cases} \phi_+(x) & x \ge 0 \\ \phi_-(x) & x < 0 \end{cases} \qquad (.10)$$

with $\phi_\pm := \phi|_{\mathbb{R}^\pm}$ smooth and $\phi(0) = \phi_-(0) = \phi_+(0)$. $\phi_\pm$ are smooth, whence they have a Taylor expansion given by

$$\phi_\pm(x) = \phi(0) + \partial\phi_\pm(0)x + \int_0^x \partial^2\phi_\pm(t)(x-t)dt \qquad (.11)$$

for $x \in \mathbb{R}^\pm$. Observe that we can write (.11) equivalently, given $C > 0$, as

$$\phi_\pm(x) = \phi(0) \pm \partial\phi_\pm(0)\,\mathrm{ReLU}(\pm x) + \int_0^C \partial^2\phi_\pm(\pm t)\,\mathrm{ReLU}(\pm x - t)dt. \qquad (.12)$$

for all $x \in \mathbb{R}^\pm$ with $\|x\| \le C$ by using Lemma 2.3. Note that we have two ReLU terms instead of four ReLU terms as implied by Lemma 2.3. These two ReLU terms dropped due to the sign of x. Denote the versions of (.12) by $\tilde{\phi}_\pm$. We can extend construction $\tilde{\phi}_\pm$ to $[-C, C]$, which we will give us

$$\tilde{\phi}_\pm(x) = \begin{cases} \phi_\pm(x) & x \in \mathbb{R}^\pm \\ \phi(0) & x \in \mathbb{R}^\mp \end{cases} \qquad (.13)$$

for all $x \in [-C, C]$. This implies that

$$\phi(x) = \tilde{\phi}_+(x) + \tilde{\phi}_-(x) - \phi(0) \qquad (.14)$$

for all $x \in [-C, C]$, and provides us with the expression we need to construct the necessary push-forwards.

Let $\mu \in \mathbb{G}_{\phi,f}$ for $f \in \mathcal{B}_\phi$. Observe that

$$f(x) = \int_{\mathbb{R}^{d+1}} \phi(\langle x|w \rangle + b)d\mu(w, b) = \int_{\mathbb{R}^{d+1}} \mathrm{ReLU}(\langle x|w \rangle + b)d\nu(w, b), \qquad (.15)$$

where the measure $\nu = \sum_{i=1}^5 \nu_i$ is the sum of measures formed from the measures

$$\nu_1 = \phi(0)\mu(\Omega)\delta_{(0,1)} \qquad (.16)$$

$$\nu_2 = \partial^1\phi(0)_+\mu \tag{.17}$$

$$\nu_3 = -\partial^1\phi(0)_-\Theta^0_\#\mu \tag{.18}$$

$$\nu_4(A) = \int_A \int_0^1 \theta_{w,b}D^2\phi_+(\theta_{w,b}u)d\Theta^1_\#(\mu\otimes\lambda)((w,b),u) \tag{.19}$$

$$\nu_5(A) = \int_A \int_0^1 \theta_{w,b}D^2\phi_-(-\theta_{w,b}u)d\Theta^2_\#(\mu\otimes\lambda)((w,b),u) \tag{.20}$$

$$\tag{.21}$$

for the Borel sets $A \subseteq \Omega$ using the push-forward maps

$$
\begin{aligned}
\Theta^0 &: \mathbb{R}^{d+} \to \mathbb{R}^{d+1}, \ (w,b) \to (-w,-b)\\
\Theta^1 &: \mathbb{R}^{d+1}\times[0,1] \to \mathbb{R}^{d+1}, \ ((w,b),u) \to (w,b-\theta_{w,b}u-y)\\
\Theta^2 &: \mathbb{R}^{d+1}\times[0,1] \to \mathbb{R}^{d+1}, \ ((w,b),u) \to (-w,-b-\theta_{w,b}u+y)
\end{aligned}
\tag{.22}
$$

where λ is the Lebesgue measure on $[0,1]$. Hence, $\rho \in \mathbb{G}_{\mathrm{ReLU},f}$. Furthermore,

$$
\begin{aligned}
\|f\|_{\mathcal{B}_{\mathrm{ReLU}}} &\le \int_{\mathbb{R}^{d+1}} (\|w\|_{\ell^1}+|b|)d|\nu|(w,b)\\
&\le \sum_{i=1}^{5}\int_{\mathbb{R}^{d+1}} (\|w\|_{\ell^1}+|b|)d|\nu_i|(w,b)\\
&\le (|\phi(0)|+|\partial^1\phi_+(0)|+|\partial^1\phi_-(0)|)\int_{\mathbb{R}^{d+1}}(1+\|w\|_{\ell^1}+|b|)d|\mu|(w,b)\\
&\quad + \int_{\mathbb{R}^{d+2}}\int_0^1 \theta_{w,b}|\partial^2\phi_+(\theta_{w,b}u)|(1+\|w\|_{\ell^1}+|b|+\theta_{w,b}|u|)dud|\mu|(w,b)\\
&\quad + \int_{\mathbb{R}^{d+1}}\int_{-1}^0 \theta_{w,b}|\partial^2\phi_-(\theta_{w,b}u)|(1+\|w\|_{\ell^1}+|b|+\theta_{w,b}|u|)dud|\mu|(w,b)\\
&\le \left(|\phi(0)|+|\partial^1\phi_+(0)|+|\partial^1\phi_-(0)|\right)\int_{\mathbb{R}^{d+2}}(1+\|w\|_{\ell^1}+|b|)d|\mu|(w,b)\\
&\quad + \sup_{(w,b)\in\mathbb{R}^{d+1}}\int_0^1 \theta_{w,b}|\partial^2\phi_+(\theta_{w,b}u)|(1+|u|)du\int_{\mathbb{R}^{d+2}}(1+\|w\|_{\ell^1}+|b|)d|\mu|(w,b)\\
&\quad + \sup_{(w,b)\in\mathbb{R}^{d+1}}\int_{-1}^0 \theta_{w,b}|\partial^2\phi_-(\theta_{w,b}u)|(1+|u|)du\int_{\mathbb{R}^{d+2}}(1+\|w\|_{\ell^1}+|b|)d|\mu|(w,b)\\
&\le \left(|\phi(0)|+|\partial^1\phi_+(0)|+|\partial^1\phi_-(0)|\right)\int_{\mathbb{R}^{d+2}}(1+\|w\|_{\ell^1}+|b|)d|\mu|(w,b)\\
&\quad + 2\sup_{(w,b)\in\mathbb{R}^{d+1}}\int_0^{\theta_{w,b}}|\partial^2\phi_+(z)|dz\int_{\mathbb{R}^{d+2}}(1+\|w\|_{\ell^1}+|b|)d|\mu|(w,b)\\
&\quad + 2\sup_{(w,b)\in\mathbb{R}^{d+1}}\int_{-\theta_{w,b}}^0|\partial^2\phi_-(z|dz\int_{\mathbb{R}^{d+2}}(1+\|w\|_{\ell^1}+|b|)d|\mu|(w,b)\\
&= \Bigg(|\phi(0)|+|\partial^1\phi_+(0)|+|\partial^1\phi_-(0)|\\
&\quad + 2\int_0^\infty|\partial^2\phi_+(z)|dz+2\int_{-\infty}^0|\partial^2\phi_-(z|dz\Bigg)\int_{\mathbb{R}^{d+2}}(1+\|w\|_{\ell^1}+|b|)d|\mu|(w,b)
\end{aligned}
$$

Taking the infimum over $\mu \in \mathbb{G}_{\phi,f}$ gives (.9). $\hspace{2cm}$ *Q.E.D.*