Journal of Machine Learning Research 25 (2024) 1-49 Submitted 11/22; Revised 8/23; Published 1/24

Optimal Bump Functions for Shallow ReLU networks: Weight Decay, Depth Separation, Curse of Dimensionality

Stephan Wojtowytsch S.WOJ@PITT.EDU
University of Pittsburgh
Department of Mathematics
Thackeray Hall
Pittsburgh, PA 15221, USA

Editor: Joan Bruna

Abstract

In this note, we study how neural networks with a single hidden layer and ReLU activation interpolate data drawn from a radially symmetric distribution with target labels 1 at the origin and 0 outside the unit ball, if no labels are known inside the unit ball. With weight decay regularization and in the infinite neuron, infinite data limit, we prove that a unique radially symmetric minimizer exists, whose average parameters and Lipschitz constant grow as d and $\sqrt{d}$ respectively.

We furthermore show that the average weight variable grows exponentially in d if the label 1 is imposed on a ball of radius ε rather than just at the origin. By comparison, a neural networks with two hidden layers can approximate the target function without encountering the curse of dimensionality.

Keywords: Deep learning, depth separation, Barron space, Radon-BV, compact support, mollifier, weight decay, minimum norm solution, symmetry learning, explicit regularization, curse of dimensionality, radial symmetry.

1. Introduction

Neural networks have revolutionized fields from computer vision (Krizhevsky et al., 2012) to natural language processing (Vaswani et al., 2017). They are the driving force behind AIs which play strategy games at superhuman levels of proficiency (Silver et al., 2016, 2017), facilitated major advances in scientific problems such as protein folding (Tunyasuvunakool et al., 2021; Jumper et al., 2021), and have been used for computer-assisted proofs in applied mathematics by Wang et al. (2022). While empirical evidence indicates that they often generalize well to previously unseen data when trained appropriately, there is little rigorous understanding of how neural networks interpolate a function between known data points.

In this article, we provide insight in the simple setting of infinitely wide ReLU networks with a single hidden layer and data which are drawn from a radially symmetric distribution on a Euclidean space $\mathbb{R}^d$. The target function f^* satisfies $f^*(0) = 1$ and $f^*(x) = 0$ for $|x| \geq 1$, where $|\cdot|$ denotes the Euclidean norm on $\mathbb{R}^d$. We consider a loss functional composed of an ℓ^2-error and a weight decay regularizer. Despite the fact that neural networks with a single hidden layer cannot represent compactly supported target functions exactly (He

et al., 2018; Lu, 2021), there are such functions which can be approximated efficiently even in high dimension. Here, we construct optimal infinitely wide networks, and show that the weight decay regularizer grows only linearly in the dimension d of the data space, improving on the quadratic upper bound established by Ongie et al. (2019).

While highly idealized, this setting allows us to study several important aspects of neural network models:

1. **Learning symmetries.** The target function has two important symmetries:
   - f^* is radially symmetric on $\{0\} \cup (\mathbb{R}^d \setminus B_1(0))$. While it is impossible to fit this symmetry exactly by finite networks, it can be attained asymptotically for highly overparametrized networks. More precisely, one could ask whether regularized risk minimization leads to symmetry learning. While we show that a unique radially symmetric solution exists, it remains open whether other solutions exist which do not exhibit radial symmetry.
   - $0 \leq f^* \leq 1$ everywhere where we prescribe a value, i.e. everywhere on $\{0\} \cup (\mathbb{R}^d \setminus B_1(0))$. Unlike linear models, which necessarily output negative data even if all training data labels are positive, but not constant, neural networks have the capacity to respect this constraint. We show that the unique radial 'minimum norm interpolant' of this prescribed data remains in $[0, 1]$ *everywhere* on $\mathbb{R}^d$.
2. **Fitting random or perturbed data.** It is known that overparametrized neural networks can fit random data, but due to the great generality of the result, the network weights may be prohibitively large for given data. The compactly supported bump functions f_d^* can be used to obtain an upper bound on necessary increase in the average size of weights if the labels y_i at points x_i are perturbed as $\tilde{y}_i = y_i + \varepsilon_i$. This is of interest in particular if the true labels y_i are generated by a target function f^* which can be approximated well by shallow neural networks with at most moderately large weights.
3. **Depth separation and curse of dimensionality.** We prove two complimentary results:
   - In dimension d, there exists an infinitely wide ReLU network with one hidden layer f_d^* with weight decay regularizer $\sim d$ such that $f_d^*(0) = 1$ and $f_d^*(x) = 0$ if $|x| \geq 1$.
   - If $f_{d,\varepsilon}^*$ is an infinitely wide ReLU network with one hidden layer such that $f_{d,\varepsilon}^*(x) = 1$ for $|x| \leq \varepsilon$ and $f_{d,\varepsilon}^*(x) = 0$ if $|x| \geq 1$, then the average parameter of f_d^* grows at least exponentially in magnitude as $\varepsilon^2 d^{1/2}(1-\varepsilon^2)^{-\frac{d+1}{2}}$ in the dimension d of the data space.

   The curse of dimensionality can be avoided in the second situation by using a neural network with two hidden layers, for which the average square weight only grows as $\sim d^{1/3}(1-\varepsilon)^{-1}$. This separation is perhaps not surprising – for instance $x \mapsto \max\{0, 1 - \|x\|_{\ell^1}\}$ is a compactly supported function which can be represented exactly by a ReLU network with two hidden layers. Neural networks with two hidden layers are more expressive in this fashion. See also (He et al., 2018) for a more detailed analysis.

4. **Effect of regularization.** We consider the Barron norm of a neural network, a measure of complexity based on the average size of the network weights see (3) and (4) for details. The expression (4) is a generalization of the weight decay (or Tikhonov) regularizer which is often added to loss functions in learning applications. Weight decay regularization is often taken as a proxy for controlling the Lipschitz constant of a neural network, as it can be computed more easily. In this highly symmetric setting, we can compare two optimal solutions:

   (a) The data is fitted optimally by the function $\hat{f}_d(x) = \max\{1 - |x|, 0\}$, which attains the minimal Lipschitz constant 1. The function cannot be represented by a ReLU network with a single hidden layer and finite weights, even in the infinite width limit (E and Wojtowytsch, 2020a, Example 5.19). It can be represented by a neural network with two infinitely wide hidden layers and weight decay $\sim \sqrt{d}$.

   (b) The Barron norm/weight decay regularizer of the optimal two-layer ReLU network f_d^* grows like d, while its Lipschitz constant grows like $\sqrt{d}$.

5. **Highly localized peaks.** The target function can be seen as the prototypical example of learning functions which take values $y_1, \dots, y_N$ at isolated points $x_1, \dots, x_N$ which are separated as 'islands' in a 'sea' of points $x_{N+1}, \dots, x_M$ with labels $y_{N+1} = \dots = y_M = 0$.

6. **Mollification.** The infinitely wide neural networks constructed in this note can be used to establish approximation rates in function spaces for shallow neural networks by mollification, if the mollification width ε is optimized to balance the competition between approximation of the target function by the infinitely wide network and approximation of the infinitely wide network by finite neural networks.

To the best of our knowledge, this is the first time that an optimal solution for fitting data by neural networks has been computed in dimension $d > 1$. For technical reasons, we focus on the case that d is odd. The optimal radial solution can be written as a finite sum

$$f_d(x) = \sum_{i=0}^{n+1} \mu_i \fint_{S^{d-1}} \sigma\big(\nu^T x - b_i\big) \, \mathrm{d}\mathcal{H}^{d-1}(\nu), \qquad n = \frac{d-1}{2}, \quad 0 = b_0 < \dots < b_{n+1} = 1$$

for some coefficients $\mu_i \in \mathbb{R}$ satisfying $\sum_{i=0}^{n} |\mu_i| = \gamma_n \sim 3.7d$.

The article is organized as follows. In the remainder of the Introduction, we briefly review the context of this work in the literature and the notation we will use throughout the article. In Section 2, we give a brief introduction to the function spaces associated to two-layer ReLU networks with a weight decay regularizer (Barron or Radon BV spaces). Sections 3 and 4 are dedicated to the statement and proof of our main results respectively. Applications of our results can be found in Section 5. Numerical approximations of the optimal solutions f_d^* can be found in Section 6. We conclude the article with a brief summary and list of open problems in Section 7.

Further numerical experiments can be found in Appendix A. Some proofs from the main part of the article are postponed to Appendix B, while proofs of results which are known in similar form are postponed to Appendix C. Slight extensions of the main results can be found in Appendix D.

1.1 Previous Work

The complexity of a neural network is often measured by the number of its non-zero coefficients (weights) (Louizos et al., 2017; Srinivas et al., 2017; Gribonval et al., 2022) or by a measure of their magnitude. From a practical perspective, both are crucial pieces of information: a neural network with an excessive number of non-zero connections is expensive to store and evaluate, while a network with very large coefficients is likely to depend on subtle cancellations at training data points and unlikely to generalize well to unseen data.

Barron (1993) realized that a large class of functions $\mathcal{F}$ can be approximated efficiently by neural networks with a single hidden layer and any sigmoidal activation function while keeping the outer layer coefficients bounded. The function class is defined in terms of a spectral criterion and diverse enough that any linear method of approximation must face the curse of dimensionality in it. More precisely, the Kolmogorov width of the unit ball $B_1 = \{f \in \mathcal{F} : \|f\|_{\mathcal{F}} \leq 1\}$ in $\mathcal{F}$ decays as

$$\sup_{dim(V)=m} \text{dist}_{L^2\left((0,1)^d\right)}\left(\mathcal{F}, V\right) = \sup_{dim(V)=m} \sup_{f\in B_1} \min_{v\in V} \|f - v\|_{L^2\left((0,1)^d\right)} \geq \frac{\bar{c}}{d}\, m^{-1/d}$$

for a universal constant $\bar{c}$. Intuitively, this means that for any m-dimensional function space V, there exists $f \in \mathcal{F}$ such that f can only be approximated to accuracy $\sim \|f\|_{\mathcal{F}} m^{-1/d}$ by elements of V . For large d, the approximation error decays extremely slowly as m increases. The function class $\mathcal{F}$ is essentially tailored to neural networks with complex exponential activation, and the main result can be understood as a statement on what we call Barron space associated to the Heaviside activation function instead of ReLU activation.

Subsequently, function approximation by ReLU networks with a single hidden layer and bounded coefficients in both layers was studied by Bach (2017); E et al. (2019b,d); E and Wojtowytsch (2020a). Optimal rates of approximation were obtained by Siegel and Xu (2019, 2021b). A spectral criterion for this scenario in terms of the Fourier transform was developed by Klusowski and Barron (2018), and a sharp criterion in terms of the Radon transform by Ongie et al. (2019); Parhi and Nowak (2021). A detailed study of Fourier-like criteria in this context is given by Caragea et al. (2020).

The norm in these function spaces is related to the popular explicit 'weight decay' regularizer (the ℓ^2-norm of the network weights). It retains significance in the context of implicit regularization, as Chizat and Bach (2020) showed that infinitely wide two-layer ReLU networks converge to minimum norm/maximum margin classifiers with respect to the weight decay norm, when trained by a gradient flow optimizer for binary classification with logistic loss.

While the structure of the function spaces has been studied and many of their functional analytic properties are understood (E and Wojtowytsch, 2020a; Parhi and Nowak, 2021; Siegel and Xu, 2021b,a), explicit examples remain rare. Spectral criteria have been used to show that functions in certain smoothness classes can be expressed as infinitely wide two-layer networks with finite weight-decay norm. E and Wojtowytsch (2022) construct a maximum margin classifier in a simple one-dimensional scenario. A structure theorem is given by E and Wojtowytsch (2020a) to easily demonstrate that certain functions cannot be expressed this way. Closest to the present work are those of Hanin (2021) and Boursier and Flammarion (2023), where the minimum norm interpolants of a finite one-dimensional

data set are studied in slightly different settings, depending on whether the magnitude of bias variables is penalized or not.

Much of the work on ReLU-activated two-layer networks makes heavy use of the homogeneity of the activation function. Two-layer neural networks with arbitrary activation are studied e.g. in Siegel and Xu (2020); Li et al. (2020). Partial (and different) extensions to deeper neural networks can be found e.g. by Parhi and Nowak (2022); E and Wojtowytsch (2020b), while residual neural networks of continuous depth ('neural ODEs') have been studied from this perspective by E et al. (2019a,d,c).

1.2 Notation

We denote by $⨍_A$ the average integral over a set A which has finite measure for a measure μ, i.e. $⨍_A f(x)\,\mathrm{d}\mu_x = \frac{1}{\mu(A)}\int_A f(x)\,\mathrm{d}\mu_x$. By $\mathrm{d}\mu_x$ we mean that we integrate with respect to the (signed) measure μ in the variable x. In this article, μ will always be a measure (often signed), while ν denotes the exterior normal vector field on a sphere.

The natural $d-1$-dimensional area (Hausdorff) measure is denoted by $\mathcal{H}^{d-1}$. In this article, it will always refer to the (unnormalized) uniform distribution on a $d-1$-dimensional sphere.

The total variation norm of a measure μ on a measurable space X is defined as $\|\mu\|_{TV} = \mu_+(X) + \mu_-(X)$, where μ_+, μ_- is the Hahn decomposition of the signed measure μ.

In the following, g is always going to be a function of one variable and f is going to be a radially symmetric function on $\mathbb{R}^d$. By an abuse of notation, we will also consider $f : [0,\infty) \to \mathbb{R}$ defined by $f(r) = f(r \cdot e_1)$. We denote by

$$c_d = \frac{|S^{d-2}|}{|S^{d-1}|} = \frac{1}{\int_{-1}^1 (1-s^2)^{\frac{d-3}{2}}\,\mathrm{d}s}$$

a quotient related to the area of hyperspheres in dimension d and $d-1$, and by γ_n a constant related to the approximability of the function $\sqrt{s}$ by polynomials of degree at most n in $L^\infty(0,1)$, which also relates to the minimal value of the weight decay regularizer for fitting data as above. The precise definition is given in Lemma 12.

The variables d and n are always related by $n = \frac{d-1}{2}$, i.e. $d = 2n+1$.

2. Weight Decay and Barron Spaces

In this section, we briefly review the theory of infinitely wide ReLU networks with a single hidden layer. Function spaces for this setting have been studied under the name $\mathcal{F}_1$ by Bach (2017), Barron space by E et al. (2020, 2019d,c,b), Radon-BV by Parhi and Nowak (2022, 2021) and the convex hull of the ReLU dictionary or the variation space of the ReLU dictionary by Siegel and Xu (2021a,c). In this note, we refer to them as Barron spaces in reference to the seminal work of Barron (1993). Some results presented below are extensions of known results to the case where we consider a Barron semi-norm rather than the full Barron norm, corresponding to a weight decay regularizer which does not control the magnitude of the biases.

A neural network with a single hidden layer and $m \in \mathbb{N}$ neurons can be represented as

$$f_m(x) = \sum_{i=1}^{m} a_i \, \sigma(w_i^T x + b_i) \qquad \text{or} \qquad f_m(x) = \frac{1}{m} \sum_{i=1}^{m} a_i \, \sigma(w_i^T x + b_i) \tag{1}$$

where $(a_i, w_i, b_i) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$ are the *weights* of the neural network. For networks in which the size of the weights is controlled, this representation can be generalized to

$$f_\mu(x) = \int_{\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}} a \, \sigma(w^T x + b) \, \mathrm{d}\mu_{(a,w,b)} \qquad \text{or } f_\pi(x) = \int_{\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}} a \, \sigma(w^T x + b) \, \mathrm{d}\pi_{(a,w,b)} \tag{2}$$

where μ is a measure on $\mathbb{R}^{d+2}$ and π is a probability measure on $\mathbb{R}^{d+2}$. More generally, due to the symmetry $a\sigma(w^T x + b) = \lambda\big((\lambda^{-1}a)\,\sigma(w^T x + b)\big)$ for $\lambda \neq 0$, μ can be taken to be a signed measure. Finite networks are contained in the general setting by setting

$$\mu_m = \sum_{i=1}^{m} \lambda_i \, \delta_{(\lambda_i^{-1} a_i, w_i, b_i)} \quad \text{and} \quad \pi_m = \frac{1}{m} \sum_{i=1}^{m} \delta_{(a_i,w_i,b_i)}$$

respectively, where the parameters $\lambda_i \neq 0$ can be chosen freely for a convenient representation. The integral is guaranteed to converge if the *Barron norm*

$$\|f\|_{\mathcal{B}} = \inf_{\pi} \left\{ \int_{\mathbb{R}^{d+2}} |a| \cdot \big\{|w| + |b|\big\} \, \mathrm{d}\pi : f \equiv f_\pi \right\} = \inf_{\mu} \left\{ \int_{\mathbb{R}^{d+2}} |a| \cdot \big\{|w| + |b|\big\} \, \mathrm{d}|\mu| : f \equiv f_\mu \right\}. \tag{3}$$

is finite, where $|\mu| = \mu^+ + \mu^-$ denotes the total variation measure of the signed measure $\mu = \mu^+ - \mu^-$. The infimum must be taken since the representation of a function in this fashion is highly non-unique (E and Wojtowytsch, 2020a, Section 2.1). The two representations of the norm coincide by (E and Wojtowytsch, 2020a, Section 2.4).

The norm in the parameter variable w is chosen dual to the norm in the data variable x such that the inequality $|w^T x| \leq |w| \cdot |x|$ holds. In particular, if distances in the data domain are measured in the ℓ^p-sense for $p \in [1, \infty]$, then distances in the parameter domain are measured in the ℓ^q-sense for $q = \frac{p}{p-1}$. For compatibility with radial symmetry, we focus on the case $p = \frac{p}{p-1} = 2$ in this note.

We refer to the space $\{f : \|f\|_{\mathcal{B}} < \infty\}$ as Barron space $\mathcal{B}$, or at times $\mathcal{B}(\mathbb{R}^d)$ to indicate dependence on dimension.

Due to the control over the bias, the Barron norm as defined by E et al. (2019d); E and Wojtowytsch (2020a) is not invariant under translations in the data space, i.e. the functions f and $f(\cdot + \bar{x})$ generally have a different norm for $\bar{x} \neq 0$. By contrast, the following *Barron semi-norm* is translation invariant and has useful properties which suffice in many applications:

$$[f]_{\mathcal{B}} = \inf_{\pi} \left\{ \frac{1}{2} \int_{\mathbb{R}^{d+2}} |a|^2 + |w|^2 \, \mathrm{d}\pi : f \equiv f_\pi \right\} = \inf_{\mu} \left\{ \frac{1}{2} \int_{\mathbb{R}^{d+2}} |a|^2 + |w|^2 \, \mathrm{d}|\mu| : f \equiv f_\mu \right\}. \tag{4}$$

We will address the convergence of the integrals in (2) without control over b in Proposition 1. This is more in line with the approach of Ongie et al. (2019); Parhi and Nowak (2021),

where the magnitude of the bias is also not controlled. We opt for controlling $|a|^2 + |w|^2$ rather than $|a| \cdot |w|$ for convenience, but note that the classical Barron norm could be defined in this fashion, too. The key observation is that the ReLU activation function $\sigma(z) = \max\{z, 0\}$ is positively one-homogeneous, i.e. $\sigma(\lambda z) = \lambda\sigma(z)$ for all $\lambda > 0$. In particular

$$a\sigma(w^T x + b) = a\sqrt{\frac{|w|}{|a|}}\,\sigma\left(\sqrt{\frac{|a|}{|w|}}w^T x + \sqrt{\frac{|a|}{|w|}}b\right),$$

i.e. we may normalize neurons (a_i, w_i, b_i) to

$$a_i' = a_i\sqrt{\frac{|w_i|}{|a_i|}}, \quad w_i' = \sqrt{\frac{|a_i|}{|w_i|}}\,w_i \qquad \text{s.t. } |a_i'|^2 = |w_i'|^2 = |a_i|\,|w_i|$$

without changing the output of the neural network. In particular $|a|^2 + |w|^2 = 2|a|\,|w|$, indicating that we could define the Barron norm in the analogous fashion by squares. Indeed, in the infinite limit it is even possible to assume that π is supported on the set $|a| = |w| = \sqrt{[f]_{\mathcal{B}}}$. For a more technically rigorous discussion, see e.g. E and Wojtowytsch (2020a).

By a slight abuse of terminology, we will also refer to $\mathcal{B}_0$ as Barron space from now on and to elements of $\mathcal{B}_0$ as Barron functions. If the two are to be distinguished, we call $\mathcal{B}_0$ the *homogeneous* Barron space. Unlike the full Barron norm, the Barron semi-norm is translation invariant and leads to spatial homogeneity. We briefly note the following properties, which relate the Barron semi-norm and more well-established quantities.

Proposition 1. *1. If the integral in* (4) *is finite for* π*, then the integral defining* f_π *in* (2) *exists for all* $x \in \mathbb{R}^d$ *if and only if it exists for* $x = 0$*. It may then be re-cast as*

$$f_\pi(x) = f_\pi(0) + \int_{\mathbb{R}^{d+2}} a\left[\sigma(w^T x + b) - \sigma(b)\right] \mathrm{d}\pi.$$

This expression always converges if the integral in (4) *is finite. The integral exists as a Bochner integral with values in* $C^0(K)$ *for compact* $K \subseteq \mathbb{R}^d$ *or* $L^p(\mathbb{P})$ *for a probability distribution* $\mathbb{P}$ *on* $\mathbb{R}^{d+2}$ *with finite* p*-th moments.*

2. $[f]_{\mathcal{B}}$ *is a norm on the homogeneous Barron space* $V_0 = \{f \in C^0(\mathbb{R}^d) : f(0) = 0, [f]_{\mathcal{B}} < \infty\}$*, which makes* V_0 *a Banach space. Compared to classical Barron spaces,* $V_0 \not\subseteq \mathcal{B}(\mathbb{R}^d)$*.*

3. $[f]_{\mathcal{B}} \leq \|f\|_{\mathcal{B}}$*.*

4. If $f \in \mathcal{B}$*, then* f *is Lipschitz-continuous and the Lipschitz-constant of* f *satisfies* $[f]_{Lip} \leq [f]_{\mathcal{B}}$*.*

All statements could be given in terms of a general signed measure μ instead of π. The proof, along with other proofs from this section, can be found in Appendix C.

Functions in Barron spaces are defined by means of an explicit representation formula. Paradoxically, this explicit characterization often makes it difficult to verify whether a given

function is in Barron space. A more abstract framework was created by Ongie et al. (2019) by the means of the Radon transform, based on the observation that

$$\Delta\left(\sum_{i=1}^m a_i\sigma(w_i^T\cdot+b_i)\right)=\sum_{i=1}^m a_i|w_i|\cdot\mathcal{H}^{d-1}|_{W_i}$$
$$D^2\left(\sum_{i=1}^m a_i\sigma(w_i^T\cdot+b_i)\right)=\sum_{i=1}^m a_i|w_i|\cdot\frac{w_i}{|w_i|}\otimes\frac{w_i}{|w_i|}\cdot\mathcal{H}^{d-1}|_{W_i},$$

i.e. the second spatial derivatives of a ReLU network with one hidden layer are superpositions of measures concentrated on the hyperplanes $W_i=\{x : w_i^Tx+b_i=0\}$. This allows for a characterization of Barron spaces in terms of second derivatives. The Radon transform is used as a technical tool in order to dualize from hyperplanes to points. This convenient characterization allows the construction of some examples of functions in Barron space.

Example 1. *1. Assume that f is a Lipschitz-continuous function and that the (possibly non-integer) power $(-\Delta)^{(d+1)/2}f$ of the Laplacian in the distributional sense exists as a measure. Then*

$$[f]_{\mathcal{B}(\mathbb{R}^d)}\leq\frac{1}{2^{d-1}\pi^{d/2-1}\Gamma(d/2)}\|(-\Delta)^{(d+1)/2}f\|_{TV},$$

where $\|\cdot\|_{TV}$ denotes the total variation norm of Δf (Ongie et al., 2019, Proposition 3).

2. If d is odd, the power of the Laplacian is integer. In particular, if f belongs to the Sobolev space $W^{d+1,1}(\mathbb{R}^d)\subseteq C^{d+1}(\mathbb{R}^d)$ of functions whose first $d+1$ (weak) partial derivatives are L^1-integrable, then $f\in\mathcal{B}(\mathbb{R}^d)$ and

$$[f]_{\mathcal{B}}\leq c_d\|f\|_{W^{d+1,1}}.$$

for some constant $c_d>0$, which depends on the exact choice of the norm on $W^{d+1,1}$. In particular $C_c^\infty(\mathbb{R}^d)\subseteq\mathcal{B}(\mathbb{R}^d)$ (Ongie et al., 2019, Corollary 1).

3. If $d\geq 3$ is an odd integer and $f_{d,k}:\mathbb{R}^d\to\mathbb{R}$ is the radial bump function given by

$$f_{d,k}(x)=\begin{cases}\left(1-|x|^2\right)^k & |x|\leq 1\\ 0 & else\end{cases},$$

then $f_{d,k}\in\mathcal{B}_0(\mathbb{R}^d)$ if $k\geq\frac{d+1}{2}$. For $k_d=\frac{d+1}{2}+2$, the norm bound $[f_{d,k_d}]_{\mathcal{B}(\mathbb{R}^d)}\leq 2d(d+5)$ holds according to (Ongie et al., 2019, Example 3).

Ongie et al. (2019) also claim a stronger version of the statement, including an if and only if condition for k and a comparable lower bound for $[f_{d,k_d}]_{\mathcal{B}(\mathbb{R}^d)}$. Those claims are based on an error in the proof of (Ongie et al., 2019, Proposition 15), where the erroneous claim is made that if $\int_{\mathbb{R}^d}|\phi|\,\mathrm{d}x=1$, then the integral of ϕ over any hyperplane $\int_H|\phi|\,\mathrm{d}\mathcal{H}^{d-1}$ is bounded from above by 1.

Based on the same intuition, we point out two observations. The first demonstrates that the singular set Σ of a Barron function (i.e. the set where the functions is not differentiable) is 'straight' and lower dimensional. This is a stronger version of Rademacher's theorem, which states that the singular set of a Lipschitz function is Lebesgue null, in the context of Barron spaces. The following statement has the stronger implication that Σ is contained in a countable union of affine subspaces of $\mathbb{R}^d$ and therefore has Hausdorff dimension $\leq d-1$.

Proposition 2. *(E and Wojtowytsch, 2020a) Any function $f \in \mathcal{B}(\mathbb{R}^d)$ can be written as a countable sum $f = \sum_{i=0}^{\infty} f_i$ where*

1. *$f_0 \in \mathcal{B}(\mathbb{R}^d)$ is C^1-smooth,*
2. *$f_i(x) = g_i(P_i x + b_i)$ where*
   - *$P_i : \mathbb{R}^d \to \mathbb{R}^{k_i}$ is an orthogonal projection for $1 \leq k_i \leq d$ (i.e. $P_i P_i^T = I_{k\times k}$)*
   - *$g_i \in \mathcal{B}(\mathbb{R}^{k_i})$ is C^1-smooth except at $0 \in \mathbb{R}^{k_i}$.*

The fact that the singular set is straight has two immediate implications.

Corollary 3. 1. *If $f \in \mathcal{B}(\mathbb{R}^d)$ is radially symmetric, then $f \in C^1(\mathbb{R}^d \setminus \{0\})$.*

2. *If $\phi : \mathbb{R}^d \to \mathbb{R}^d$ is a diffeomorphism such that $f \in \mathcal{B}(\mathbb{R}^d) \Rightarrow f \circ \phi \in \mathcal{B}(\mathbb{R}^d)$, then ϕ is an affine linear map (E and Wojtowytsch, 2020a, Theorem 5.18).*

A brief inspection of the proof of (E and Wojtowytsch, 2020a, Theorem 5.18) reveals that Proposition 2 and Corollary 3 remain valid for $\mathcal{B}_0(\mathbb{R}^d)$. A stronger result on radial Barron functions is proved below in Lemma 10. Secondly, we recall a characterization of one-dimensional Barron spaces, which is essentially the simpler one-dimensional case of the Radon transform construction. A similar statement can also be found e.g. in (E and Wojtowytsch, 2020a, Example 4.1) and (Li et al., 2020).

Proposition 4. *$\phi \in \mathcal{B}_0(\mathbb{R})$ if and only if there exists a finite signed measure μ such that $\phi'' = \mu$, i.e. $\phi'(s) = \mu\big((-\infty, s]\big)$ for all $s \in \mathbb{R}$ such that $\mu(\{s\}) = 0$ (in particular, all but countably many). For all such ϕ and any $a \in \mathbb{R}$, we can write*

$$\phi(z) = \phi(a) + \phi'(a)\big[\sigma(x-a) - \sigma(a-x)\big] + \int_a^\infty \phi''(s)\,\sigma(z-s)\,\mathrm{d}s + \int_{-\infty}^a \phi''(s)\,\sigma(s-z)\,\mathrm{d}s.$$

Furthermore

$$[\phi]_{\mathcal{B}} \leq \|\phi''\|_{TV} + 2\inf_{a\in\mathbb{R}} \inf_{v\in\partial\phi(a)} |v|$$

where

$$\partial_a f = \operatorname{conv}\left(\left\{v \in \mathbb{R} : \exists\ x_n \to a\ \text{ s.t. } \frac{f(x_n) - f(a)}{x_n - a} \to v\right\}\right)$$

is the convex hull of the set of approximate derivatives. Conversely

$$\max\left\{\|\phi''\|_{TV}, \sup_{a\in\mathbb{R}} \inf_{v\in\partial\phi(a)} |v|\right\} \leq [\phi]_{\mathcal{B}}.$$

We believe that the upper bound is, in fact an identity. We now recall a property of Barron spaces $\mathcal{B}_0$.

Proposition 5 (Direct approximation theorem)**.** *For every $f \in \mathcal{B}_0$ and every probability measure $\mathbb{P}$ on $\mathbb{R}^d$ there exists f_m as in* (1) *and $c > 0$ such that*

$$\|f - f_m - c\|^2_{L^2(\mathbb{P})} \leq \frac{[f]^2_{\mathcal{B}}}{m} \max_{|\nu|\leq 1} \int_{\mathbb{R}^d} |\nu^T x|^2 \,\mathrm{d}\mathbb{P}$$

and

$$|a_i| = |w_i| = \sqrt{\frac{\|f\|_{\mathcal{B}}}{m}} \qquad \text{or } |a_i| = |w_i| = \sqrt{\|f\|_{\mathcal{B}}},$$

depending on the normalization in (1)*.*

For the sake of completeness, we sketch a probablistic proof in Appendix C. This formulation of the direct approximation theorem improves on known results in two major ways:

1. The dependence on the data distribution $\mathbb{P}$ is only through the 'projected second moments' $M_{2,proj}(\mathbb{P}) := \max_{|w|\leq 1} \int_{\mathbb{R}^d} |w^T x|^2 \,\mathrm{d}\mathbb{P}$ rather than the full second moments $M_2(\mathbb{P}) := \int_{\mathbb{R}^d} |x|^2 \,\mathrm{d}\mathbb{P}$. It is easy to see that

   $$M_{2,proj}(\mathbb{P}) \leq M_2(\mathbb{P}) = \sum_{i=1}^d \int_{\mathbb{R}^d} |e_i^T x|^2 \,\mathrm{d}\mathbb{P} \leq d \cdot M_{2,proj}(\mathbb{P})$$

   for any probability measure $\mathbb{P}$ on $\mathbb{R}^d$, and that equality is attained for any measure $\mathbb{P}$ which is the product of d one-dimensional probability measures, e.g. a standard normal distribution. The constant in the bound may therefore be significantly smaller in high dimension.

2. The bound depends on the Barron semi-norm, but not the full Barron norm.

While the constants are improved in this formulation compared to e.g. E et al. (2019b,d); E and Wojtowytsch (2020a), the result is not expected to be sharp in terms of the rate which is achieved. An improvement from $m^{-1/2}$ to $m^{-1/2-3/2d}$ in the classical setting was found by Siegel and Xu (2021b) at the cost of a more involved proof.

Many of the results above are somewhat specific to ReLU activation as the proofs either use positive homogeneity or the property that $\sigma'' = \delta$. Both are shared by leaky ReLU activation.

Remark 6. *Consider the leaky ReLU activation function $\sigma_\varepsilon(z) = \max\{\varepsilon z, z\}$ for $\varepsilon \in (0,1)$ in addition to the classical ReLU activation $\sigma = \sigma_0$. Since*

$$\sigma_\varepsilon(z) = \sigma(z) - \varepsilon\,\sigma(-z) \qquad \text{and} \qquad \sigma(z) = \frac{1}{1-\varepsilon^2}\sigma_\varepsilon(z) + \frac{\varepsilon}{1-\varepsilon^2}\sigma_\varepsilon(-z), \tag{5}$$

any function which can be represented as a superposition of ReLUs can be represented as a superposition of leaky ReLUs and vice versa. The entire construction of Barron space goes

through as above, leading to two semi-norms $[\cdot]_{\mathcal{B}}$ *and* $[\cdot]_\varepsilon$ *on the same function class such that* $[f]_\varepsilon \leq (1+\varepsilon)[f]_{\mathcal{B}}$ *and* $[f]_{\mathcal{B}} \leq \frac{1+\varepsilon}{1-\varepsilon^2}[f]_\varepsilon = \frac{1}{1-\varepsilon}[f]_\varepsilon$ *by the explicit representation* (5). *More compactly, we write this as*

$$(1-\varepsilon)\,[f]_{\mathcal{B}} \leq [f]_\varepsilon \leq (1+\varepsilon)\,[f]_{\mathcal{B}} \qquad \forall\ f \in \mathcal{B}_0. \tag{6}$$

Using positive one-homogeneity, it can be seen that the coefficients in the representations (5) *are in fact optimal and thus that* (6) *is sharp. The norms induced on Barron space by ReLU and leaky ReLU activation are therefore equivalent, and all properties mentioned above survive if* σ *is replaced by* σ_ε.

The more subtle statements which we prove below do not survive passing to an equivalent norm. When minimizing $[f]_\varepsilon$ *under the constraints* $f(x_i) = y_i$, *the set of solutions* $\mathcal{M}_\varepsilon \subseteq \mathcal{B}_0$ *will generally depend on* $\varepsilon \in [0,1)$. *For example, consider the one-dimensional data set with two points* $(x_0, y_0) = (0,0)$ *and* $(x_1, y_1) = (1,1)$, *which is fit exactly by* σ_ε *for any* ε. *The solution* σ_ε *is norm-minimizing for* $[\cdot]_\varepsilon$, *but not for* $[\cdot]_{\mathcal{B}}$, *where the norm-inequality is sharp (and vice versa).*

The equivalence of norms estimate degenerates at $\varepsilon = 1$, *where the activation would become linear. If* $\varepsilon < 0$, *a similar construction holds unless* $\varepsilon = -1$, *where any* σ_ε*-Barron function* f *would have to satisfy* $\lim_{t\to\infty} f(tx) = \lim_{t\to-\infty} f(tx)$.

We are finally ready to state (and prove) the main results of this article rigorously.

3. Statements of Main Results

Theorem 7. *For every odd* $d \in \mathbb{N}$, *there exists a unique radial function* $f_d^* \in \mathcal{B}(\mathbb{R}^d)$ *such that*

$$f_d^* \in \operatorname*{argmin}_{f\in\mathcal{F}} [f]_{\mathcal{B}}, \qquad \mathcal{F} := \left\{ f \in C(\mathbb{R}^d) : f(0) = 1 \text{ and } f \equiv 0 \text{ on } \mathbb{R}^d \setminus B_1(0) \right\}.$$

Furthermore

1. $f_d^* \in C^{\frac{d-1}{2}}(\mathbb{R}^d \setminus \{0\})$.
2. *The radial profile* $\hat{f}_d^* : [0,\infty) \to \mathbb{R}$, $\hat{f}_d^*(r) = f_d^*(r \cdot e_1)$ *is strictly monotone decreasing in* r *in* $(0,1)$. *In particular,* $0 \leq f_d^* \leq 1$.
3. *There exists* $r_d > 0$ *such that* $\hat{f}_d^*$ *is a* linear, *strictly monotone decreasing function of* r *on* $[0, r_d]$.

As $d \to \infty$, *the norm of* f_d^* *increases linearly as*

$$\lim_{d\to\infty,\, d \text{ odd}} \frac{[f_d^*]_{\mathcal{B}(\mathbb{R}^d)}}{d} = \gamma \approx 3.6,$$

where γ *is the inverse of the Bernstein constant.*

The Bernstein constant is a quantity in classical numerical analysis and approximation theory arising when approximating the function $h(x) = |x|$ by polynomials in $L^\infty(-1,1)$,

see e.g. (Trefethen, 2019). From the proof of Theorem 7, we obtain an algorithm to compute f_d^* to arbitrary precision, which is implemented in Section 6.

The functions f_d^* are radially symmetric, compactly supported and non-negative. In particular, they can serve as mollifiers to easily prove quantitative approximation results for two-layer ReLU networks in general function classes. In an upcoming companion article (Park et al., 2023), we prove that they are achieved as (radial averages of) empirical risk minimizers with a weight decay regularizer.

Note that we do not exclude the possibility that other minimizers exist which are not radially symmetric. From direct arguments, we can only conclude that the set of minimizers is convex and invariant under coordinate rotations. The existence of at least one radially symmetric minimizer follows relatively easily, while its uniqueness is established below by construction. For *any* minimizer $\tilde{f}_d \in \operatorname{argmin}_{f\in\mathcal{F}}[f]_{\mathcal{B}}$, which may not not be radially symmetric, the radial average

$$\tilde{f}_{d,av}(x) = \fint_{SO(d)} \tilde{f}_d(Ox)\mathrm{d}H_O$$

is a radially symmetric minimizer, i.e. $\tilde{f}_{d,av} \equiv f_d^*$. Knowledge of the unique minimizer after radial averaging allows us to study optimization algorithms for implicit bias and finding global optima. This line of inquiry is pursued in upcoming work (Park et al., 2023).

We find it easier to deal with odd dimensions, as the function $(1-s^2)^{\frac{d-1}{2}}$ is a polynomial in this case. This is analogous to the observations of Ongie et al. (2019). We remark that, if $f : \mathbb{R}^D \to \mathbb{R}$ is a Barron function and $d \leq D$, then

$$\widetilde{f} : \mathbb{R}^d \to \mathbb{R}, \qquad \widetilde{f}(x) = f(x_1, \dots, x_d, 0, \dots, 0)$$

is also a Barron function and $[\widetilde{f}]_{\mathcal{B}(\mathbb{R}^d)} \leq [f]_{\mathcal{B}(\mathbb{R}^D)}$, so the limit

$$\lim_{d\to\infty} \inf \left\{ [f]_{\mathcal{B}(\mathbb{R}^d)} : f(0) = 1 \text{ and } f \equiv 0 \text{ on } \mathbb{R}^d \setminus \overline{B_1(0)} \right\} \approx 3.6$$

remains valid if even dimensions are considered, as can be seen when sandwiching an even integer d between $d-1$ and $d+1$.

Using a reflection argument, any radially symmetric Barron function $f : \mathbb{R}^d \to \mathbb{R}$ can be written as

$$f(x) = f(0) + \int_{[0,\infty)} \fint_{S^{d-1}} \sigma(\nu^T x - b)\mathrm{d}\mathcal{H}^{d-1}\,\mathrm{d}\mu_b \tag{7}$$

for some measure μ on the space of biases. In this context, Theorem 7 can be understood as a finite representer theorem, since the proof shows precisely that there exist $n+2 = \frac{d+3}{2} \in \mathbb{N}$ weights $\mu_0, \dots, \mu_{n+1}$ and biases $0 = b_0 < \dots < b_{n+1} = 1$ such that

$$f_d^*(x) = 1 + \sum_{i=0}^{n+1} \mu_i \fint_{S^{d-1}} \sigma(\nu^T x - b_i)\mathrm{d}\mathcal{H}^{d-1}.$$

Finally, we note that the methods in the proof of Theorem 7 can also be used to show the following extension.

Theorem 8. *For every $\varepsilon \in (0,1)$ and every odd $d \in \mathbb{N}$, there exists a unique radial function $f^*_{d,\varepsilon} \in \mathcal{B}(\mathbb{R}^d)$ which minimizes the Barron semi-norm in the class*

$$f^*_{d,\varepsilon} \in \operatorname*{argmin}_{f \in \mathcal{F}_\varepsilon} [f]_{\mathcal{B}}, \qquad \mathcal{F}_\varepsilon := \left\{ f \in C(\mathbb{R}^d) : f \equiv 1 \text{ on } \overline{B_\varepsilon(0)} \text{ and } f \equiv 0 \text{ on } \mathbb{R}^d \setminus B_1(0) \right\}.$$

Furthermore

1. *$f^*_{d,\varepsilon} \in C^{\frac{d-1}{2}}(\mathbb{R}^d)$.*
2. *The radial profile $\hat{f}^*_{d,\varepsilon} : [0,\infty) \to \mathbb{R}$, $\hat{f}^*_{d,\varepsilon}(r) = f^*_{d,\varepsilon}(r \cdot e_1)$ is strictly monotone decreasing on $[\varepsilon, 1]$. In particular, $0 \leq f^*_{d,\varepsilon} \leq 1$.*

In this case, the Barron norm grows exponentially in the dimension d. More precisely, there exists $D \in \mathbb{N}$ independent of $\varepsilon > 0$ such that

$$\|f^*_{d,\varepsilon}\|_{\mathcal{B}(\mathbb{R}^d)} \geq \frac{\varepsilon^2 \sqrt{d}}{(1-\varepsilon^2)^{\frac{d+1}{2}}}$$

if $d \geq D$.

We thus observe that the problem of approximating compactly supported bump functions which are constant in a neighbourhood of the origin by shallow neural networks suffers from the curse of dimensionality. We will argue below that this is not the case for ReLU networks with at least two hidden layers.

Remark 9. *Minimum norm interpolants in homogeneous spaces have been characterized by different means for general finite data sets by Savarese et al. (2019) in one dimension and by Boursier and Flammarion (2023) for the full Barron norm. Ardeshir et al. (2023) show that the minimum norm solution for fitting labels $y_i \in \{-1, 1\}$ on the vertices of a hypercube $\{x_1, \ldots, x_{2^d}\}$ are generally not ridge functions, even if the data can be fit by a ridge function.*

4. Proofs of the Main Results

We begin by stating three lemmas in this section, which are used to prove the main theorems. The proofs are given in Appendix B. By a slight abuse of notation, we denote $f(r) = f(re_1)$ for a radially symmetric function $f : \mathbb{R}^d \to \mathbb{R}$ and by f' the radial derivative of f. We first note a general result on radially symmetric Barron functions.

Lemma 10. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a radially symmetric Barron function and d odd. Then*

1. *as a function of r, f is $n := \frac{d-1}{2}$ times continuously differentiable in $\mathbb{R}^d \setminus \{0\}$. The $n+1$-th radial derivative is bounded and measurable, and the $n+2$-th radial derivative in the distributional sense is a bounded (Radon) measure.*
2. *for every $\varepsilon > 0$, there exists $D \in \mathbb{N}$ such that the Lipschitz bound*

$$[f]_{Lip} \leq \frac{1+\varepsilon}{\sqrt{2\pi d}} [f]_{\mathcal{B}_0}$$

holds for every $d \geq D$.

The following Lemma allows us to express radial symmetry and compact support for Barron functions in odd dimensions in a one-dimensional fashion. It is based on an exchange in the order of integration in (7).

Lemma 11. *Assume that $g \in \mathcal{B}_0(\mathbb{R})$ is a one-dimensional Barron function such that $g(0) = 1$, $g \equiv 0$ outside of $(-1, 1)$ and*

$$\int_{-1}^{1} g(s)\, s^{2k} \,\mathrm{d}s = 0 \qquad \text{for } k = 1, \dots, \frac{d-3}{2}. \tag{8}$$

Then the function $f : \mathbb{R}^d \to \mathbb{R}$ given by

$$f(x) = \frac{\int_{-1}^{1} (1-s^2)^{\frac{d-3}{2}}\, g(|x|s)\, \mathrm{d}s}{\int_{-1}^{1} (1-s^2)^{\frac{d-3}{2}}\, \mathrm{d}s} \tag{9}$$

satisfies the following properties:

1. $f(0) = 1$,
2. $f(x) = 0$ *if* $|x| \geq 1$,
3. f *is radially symmetric, and*
4. $[f]_{\mathcal{B}(\mathbb{R}^d)} \leq [g]_{\mathcal{B}(\mathbb{R})}$.

Conversely, if $f : \mathbb{R}^d \to \mathbb{R}$ is a radially symmetric Barron function which satisfies $f(0) = 1$ and $f \equiv 0$ on $\mathbb{R}^d \backslash B_1(0)$, then there exists g as above such that (9) holds, which is additionally an even function and satisfies $[f]_{\mathcal{B}(\mathbb{R}^d)} = [g]_{\mathcal{B}(\mathbb{R})}$.

Furthermore $f \equiv 1$ in $B_\varepsilon(0)$ if and only if $g \equiv 1$ in $(-\varepsilon, \varepsilon)$ for the even *representative of the function class g.*

We will show that such a Barron function g indeed exists for every odd $d \geq 3$ and compute the precise asymptotic growth of $[g]_{\mathcal{B}}$ as $d \to \infty$. The following Lemma is the main technical tool in our proof.

Lemma 12. *For $n \in \mathbb{N}$, set*

$$\gamma_n := \min \left\{ \|\mu\|_{TV} : \int_0^1 s \,\mathrm{d}\mu_s = 1, \quad \int_0^1 s^{2k} \,\mathrm{d}\mu_s = 0 \text{ for } 0 \leq k \leq n \right\}.$$

Then $\lim_{n\to\infty} \frac{\gamma_n}{n} = \gamma \approx 3.57$ is the inverse of the Bernstein constant. The minimum is attained by a unique measure $\mu = \sum_{i=0}^{n+1} \mu_i \delta_{s_i}$ where

1. *$0 = s_0 < s_1 < \cdots < s_{n+1} = 1$ are the $n+2$ distinct points in $[0, 1]$ at which $P(s) - s$ is extremal in $[0, 1]$, where P is the optimal even polynomial approximator of degree $\leq 2n$ for $g(s) = s$ in $L^\infty(0, 1)$.*

2. $\mu_0, \dots, \mu_{n+1} \in \mathbb{R}$ *are parameters satisfying the alternation criterion* $\mu_{i+1}\mu_i < 0$ *for* $i = 0, \dots, n$ *and the generalized Vandermonde system*

$$\begin{pmatrix} s_0 & s_1 & \dots & s_{n+1} \\ 1 & 1 & \dots & 1 \\ s_0^2 & s_1^2 & \dots & s_{n+1}^2 \\ s_0^4 & s_1^4 & \dots & s_{n+1}^4 \\ \vdots & \vdots & \ddots & \vdots \\ s_0^{2n} & s_1^{2n} & \dots & s_{n+1}^{2n} \end{pmatrix} \begin{pmatrix} \mu_0 \\ \vdots \\ \mu_{n+1} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

We are finally prepared to prove our main results. We begin with Theorem 7.

Proof Step 1. In this step, we construct g and f using Lemmas 11 and 12. Let $d \geq 3$ be an odd integer and $n = \frac{d-1}{2}$. Let $g : [0, \infty) \to \mathbb{R}$ be the unique function such that

$$g(0) = 0, \qquad \lim_{z \nearrow -1} g'(z) = 0, \qquad g'' = \mu,$$

in the distributional sense, where μ is the even reflection of the measure μ_{n+1} described in Lemma 12, i.e. $\mu(U) = \mu_{n+1}(U \cap [0, \infty)) + \mu_{n+1}(-U \cap [0, \infty))$. Note that the origin is counted twice. By construction, g is piecewise linear, $g \equiv 0$ on $(-1, 1)$ and

$$g(s) = \int_{-1}^{s} (z - s) \, \mathrm{d}\mu_z \qquad \forall\ s \geq -1.$$

In particular, $g(0) = 1$ and $g(s) = 0$ for all $s \geq 1$ due to the moment conditions

$$\int_0^1 1 \, \mathrm{d}\mu_s = 0, \qquad \int_0^1 s \, \mathrm{d}\mu_s = 1.$$

Since $g'' = \mu$ is even and $g(-1) = g(1)$, we find that g is even. Due to Proposition 4, we observe that $[g]_{\mathcal{B}} = \|\mu\|_{TV} = 2\gamma_{n+1}$. Integrating by parts twice, we realize that

$$\int_{-1}^{1} g(s)\, s^{2k} \, \mathrm{d}s = \int_{-r}^{r} g(s)\, s^{2k} \, \mathrm{d}s = \frac{1}{(2k+2)(2k+1)} \int_{-r}^{r} g''(s)\, s^{2k+2} \, \mathrm{d}s = 0 \tag{10}$$

for $r > 1$ and $k = 0, \dots, n-1 = \frac{d-3}{2}$, since $g \equiv g' \equiv 0$ in a neighbourhood of r, so the boundary terms vanish. The integration by parts is well-established for smooth functions and can be justified in the piecewise case by mollification.

In particular, g satisfies the conditions of Lemma 11 and induces an admissible radially symmetric function $f \in \mathcal{B}_0(\mathbb{R}^d)$.

Step 2. Assume for now that there exists a minimizer of the Barron semi-norm in $\mathcal{F}$. Since the Barron semi-norm is a a convex function on the convex function class $\mathcal{F}$, and since furthermore both $\mathcal{F}$ and $[\cdot]_{\mathcal{B}}$ are invariant under coordinate rotations, we note that the set of minimal semi-norm elements in $\mathcal{F}$ is both convex and rotation invariant. In particular

$$\hat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} [f]_{\mathcal{B}} \qquad \Rightarrow \quad \hat{f}_O \in \operatorname*{argmin}_{f \in \mathcal{F}} [f]_{\mathcal{B}} \quad \text{where } \hat{f}_O(x) = \int_{SO(d)} \hat{f}(Ox) \, \mathrm{d}H_O$$

is the average of $\hat{f}$ with respect to all rotations. The measure H is the Haar measure on the group $SO(d)$, i.e. the $d(d-1)/2$-dimensional Hausdorff measure induced by the Frobenius norm on the space of $d \times d$-matrices.

Thus, if a minimizer exists, then there is also a radially symmetric minimizer. In this step, we illustrate that the function $f = f_d^*$ associated to g as in Step 1 is in fact optimal. In particular, we can conclude from the proof below that a minimizer does exist.

It is easy to see that (10) is both necessary and sufficient to imply the moment conditions for g in Lemma 11. In particular

$$\inf_{f \in \mathcal{F}} [f]_{\mathcal{B}} = \inf \Bigg\{ [g]_{\mathcal{B}} : g(0) = 1,\ g \equiv 0 \text{ on } [1, \infty),\, g \text{ even and} \\ \int_0^1 g''(s)\, s^{2k} \, \mathrm{d}s = 0 \text{ for } 0 \leq k \leq \frac{d-1}{2} \Bigg\},$$

with corresponding minimizers. As $g'' = \mu$ is the unique solution to the minimization problem on the right, f is the unique radial minimizer on the left.

In particular

$$\lim_{d \to \infty, d \text{ odd}} \frac{[f_d^*]_{\mathcal{B}}}{d} = \lim_{d \to \infty, d \text{ odd}} \frac{2\gamma_{(d-1)/2}}{d} = \gamma \leq 3.6.$$

Step 3. We note that g is linear on the interval $[0, s_1]$, where s_1 is as in Lemma 12. Thus

$$f(x) = c_d \int_{-1}^1 g(|x|s)\big(1 - s^2\big)^{\frac{d-3}{2}} \, \mathrm{d}s = c_d \int_{-1}^1 \big(1 - \mu_0 |x|\, |s|\big)\big(1 - s^2\big)^{\frac{d-3}{2}} \, \mathrm{d}s \\ = 1 - |x| \, \frac{\mu_0 \int_0^1 s\big(1 - s^2\big)^{\frac{d-3}{2}} \, \mathrm{d}s}{\int_0^1 \big(1 - s^2\big)^{\frac{d-3}{2}} \, \mathrm{d}s}$$

is linear by the origin.

Step 4. In this step, we show that f is strictly decreasing in radial direction inside the unit ball. As noted in Corollary 3, the function f is C^1-smooth away at the origin. Since $f(0) = 1$ and $f(e_1) = 0$, it suffices to show that $\partial_r f(re_1) \neq 0$ for $r \in (0,1)$. We compute

$$f(re_1) = c_d \int_{-1}^1 g(rs) \left(1 - s^2\right)^{\frac{d-3}{2}} \mathrm{d}s, \qquad \partial_r f(re_1) = c_d \int_{-1}^1 g'(rs)\, s\big(1 - s^2\big)^{\frac{d-3}{2}} \, \mathrm{d}s.$$

We make the following claim: *If g is an even piecewise linear function on $[-1,1]$ with at most $n+1$ segments in $[0,1]$ and $k \leq n-1$, then the function*

$$r \mapsto \int_{-1}^1 g'(rs)\, s\big(1 - s^2\big)^k \, \mathrm{d}s$$

has at most $n - 2 - k$ zeros in $(0,1)$.

To prove the claim, start with $k = 0$. Then

$$\int_{-1}^1 g'(rs)\, s \, \mathrm{d}s = \frac{1}{r^2} \int_{-1}^1 g'(rs)\, rs\, r \, \mathrm{d}s = \frac{2}{r^2} \int_0^r g'(z)\, z \, \mathrm{d}z.$$

As g' is constant in the interval $[0, s_1]$ by the origin, $\partial_r f$ is constant (and non-zero) in $[0, s_1]$, meaning that $\partial_r f$ cannot have a zero in $[0, s_1]$. In any interval (s_i, s_{i+1}) where g' is constant, the function

$$r \mapsto \int_0^r g'(z)\, z \,\mathrm{d}z = \int_0^{s_i} g'(z)\, z \,\mathrm{d}z + \int_{s_i}^r g'(z)\, z \,\mathrm{d}z$$

is monotone, since $g'(z) \cdot z$ does not change sign. In particular:

1. There is no zero in the first interval $[s_0, s_1] = [0, s_1]$.
2. There is at most one zero in $[s_i, s_{i+1}]$.
3. The zero in the final interval $[s_n, s_{n+1}] = [s_n, 1]$ is attained at $s = 1$.

Thus there are at most $n-2$ zeros in $(0,1)$, which proves the claim for $k = 0$. Now consider $k \geq 1$. Note that

$$\int_0^1 g'(rs)\, s(1-s^2)^k \,\mathrm{d}s = r^{-(2+k)} \int_0^1 g'(rs)\, rs\big(r^2-(rs)^2\big)^k \, r \,\mathrm{d}s = r^{-(2+k)} \int_0^r g'(z)\big(r^2-z^2\big)^k \,\mathrm{d}z$$

In particular, the term on the left is zero if and only if the integral on the right is zero. If there are two points r_1, r_2 on the right where the integral vanishes (and $k \geq 1$), then by Rolle's theorem in between there exists a point $r \in (r_1, r_2)$ at which

$$0 = 2r \int_0^r g'(z)\, z\big(r^2 - z^2\big)^{k-1} \,\mathrm{d}z = r^{k-2} \int_{-1}^1 g'(rs)\, s(1-s^2)^{k-1} \,\mathrm{d}s$$

The integral also vanishes at zero, where we are integrating over the empty set. We note that for any $k \geq 0$ the integral $\int_{-1}^1 g'(rs)\, s(1-s^2)\,\mathrm{d}s$ vanishes at $r = 0$ and $r = 1$. If, for $k \geq 1$ it vanishes at N interior points, then for $k-1$ it must vanish at $N+1$ points: 0, 1, and at least once in each interval. In particular, for $k \geq 1$, there are at most $n-2-k$ interior vanishing points.

Step 5. We finally note that the Lipschitz bound follows directly from the Barron norm bound on f_d^* and Lemma 10. ∎

Example 2. *Let us consider the case $d = 3$, i.e. $n = \frac{d-1}{2} = 1$. The $n+2 = 3$ points s_0, s_1, s_2 are given by the equi-oscillating points of the best approximation of the function $f(s) = s$ on $[0,1]$ by elements of the space spanned by $\{s^0, s^2, \ldots, s^{2n}\} = \{1, s^2\}$.*

The best approximation of s by even quadratic polynomials in $L^\infty(0,1)$ is $P(s) = s^2 + \frac{1}{8}$, which attains maximal distance at $s = 0, 1/2, 1$. This can easily be verified as $P(s) - s$ is a polynomial of degree 2 inside $(0,1)$, so if $P(0) = P(1)$, then $P(s) = \alpha + \beta(s - 1/2)^2$, so the most distant points are in $\{0, 1/2, 1\}$. By Kolmogorov's equi-oscillation theorem, all three are points of largest error. It is now easy to solve for the coefficients of P.

We can find the measure $\mu = \mu_0\delta_0 + \mu_1\delta_{1/2} + \mu_2\delta_1$ for the second derivative $g'' = \mu$ by solving the linear system of moment conditions

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1/2 & 1 \\ 0 & 1/4 & 1 \end{pmatrix} \begin{pmatrix} \mu_0 \\ \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \Leftrightarrow \quad \begin{pmatrix} \mu_0 \\ \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} -3 \\ 4 \\ -1 \end{pmatrix}. \tag{11}$$

So g is the even continuous piecewise linear function satisfying $g(0) = 1$ and

$$g'(s) = \begin{cases} \mu_0 & s \in (0, 1/2) \\ \mu_0 + \mu_1 & s \in (1/2, 1) \\ \mu_0 + \mu_1 + \mu_2 & s > 1 \end{cases} \qquad \Rightarrow \qquad g(s) = \begin{cases} 1 - 3s & s \in [0, 1/2] \\ -1 + s & s \in [1/2, 1] \\ 0 & s \geq 1. \end{cases}$$

Finally, since $\frac{d-3}{2} = 0$, we find that $(1 - s^2)^{\frac{d-3}{2}} \equiv 1$ and thus

$$f(re_1) = \frac{\int_{-1}^1 g(rs)\,\mathrm{d}s}{\int_{-1}^1 1\,\mathrm{d}s} = \int_0^1 g(rs)\,\mathrm{d}s = \frac{1}{r}\int_0^r g(s)\,\mathrm{d}s = \frac{1}{r}\begin{cases} r - \frac{3}{2}r^2 & 0 \leq r \leq 1/2 \\ \frac{r^2}{2} - r + \frac{1}{2} & 1/2 \leq r \leq 1 \\ 0 & r \geq 1 \end{cases}$$

$$= \begin{cases} 1 - \frac{3}{2}r & 0 \leq r \leq 1/2 \\ \frac{r}{2} - 1 + \frac{1}{2r} & 1/2 \leq r \leq 1 \\ 0 & r \geq 1 \end{cases}.$$

In particular, we observe that $f \geq 0$ and that $f \in C^1(0, \infty)$. It is easy to see that the first derivative of f

$$f'(r) = \frac{1}{2}\left(-3 \cdot \chi_{(0,1/2]}(r) + \left[1 - \frac{1}{r^2}\right] \cdot \chi_{(1/2,1]}(r)\right)$$

is a continuous function, the second

$$f''(r) = \frac{1}{r^3} \cdot \chi_{(1/2,1)}$$

is a bounded and measurable function, and the third (distributional) derivative

$$f'''(r) = 8 \cdot \delta_{1/2} - \frac{3}{r^4} \cdot \mathcal{L}|_{(1/2,1)} - \delta_1$$

is a finite measure, where δ_x denotes a Dirac delta located at the point x and $\mathcal{L}_U$ denotes the one-dimensional Lebesgue measure of the open set U.

We now give the proof of Theorem 8.

Proof The existence of a radial minimizer $f^*_{d,\varepsilon}$ is proved as in Theorem 7. By Lemma 10, we find that $f^*_{d,\varepsilon} \in C^{\frac{d-1}{2}}(\mathbb{R}^d \setminus \{0\})$, and since $f^*_{d,\varepsilon}$ is constant in a neighbourhood of the origin, we find that $f^*_{d,\varepsilon} \in C^{\frac{d-1}{2}}(\mathbb{R}^d)$. The uniqueness follows as in Theorem 7 by considering the optimal measure μ on $[\varepsilon, 1]$ satisfying the moment conditions, using again Lemma 11. The main difference lies in the greater ability to uniformly approximate the function $f(s) = s$ by even polynomials on $[\varepsilon, 1]$ compared to $[0, 1]$.

We claim the following: *Let $\varepsilon > 0$, $n \in \mathbb{N}$ and μ_n a measure on $[\varepsilon, 1]$ such that*

$$\int_\varepsilon^1 s\,\mathrm{d}\mu_n = 1, \qquad \int_\varepsilon^1 s^{2k}\,\mathrm{d}\mu_n = 0 \qquad \forall\ k = 0, \dots, n.$$

Then for every $c < 1$ there exists $N \in \mathbb{N}$ independent of ε such that

$$\|\mu_n\| \geq c\,\frac{\varepsilon^2\sqrt{\pi n}}{(1 - \varepsilon^2)^{n+1}}$$

if $n \geq N$.

The claim is proved in Appendix B. Inserting the lower bound in (21), the statement is proved. ∎

5. Applications

5.1 Fitting Values on a Finite Data Set

Let $(x_i, y_i)_{i=1}^N$ be a finite data set in $\mathbb{R}^d \times \mathbb{R}$. For each i, define $r_i = \min_{j \neq i} |x_j - x_i|$ to be the minimal distance between the point x_i and the closest data point to it. Then

$$f(x) = \sum_{i=1}^{N} y_i \, f_d^* \left(\frac{x - x_i}{r_i} \right)$$

is a Barron function such that

1. $f(x_i) = y_i$ for all i and
2. $\|f\|_{\mathcal{B}} \leq 2\gamma_{(d+1)/2} \sum_{i=1}^N \frac{|y_i|}{r_i}$.

In most practical data sets, the minimum ℓ^2-distance between data points is lower bounded as $\Omega(1)$ or even $\Omega(\sqrt{d})$, meaning that the Barron norm only grows as $\sim dN$ or even $\sqrt{d}N$. Using the direct approximation theorem for Barron functions (Proposition 5) in $L^2(\mathbb{P}_n)$ for $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, for every $m \in \mathbb{N}$ there exists a shallow neural network f_m with m neurons (and one constant shift) such that

$$\frac{1}{n} \sum_{i=1}^{n} \left| f(x_i) - y_i \right|^2 \leq \frac{\|f\|_{\mathcal{B}}^2 \, \max_{|\nu|=1} \langle \sum_{i=1}^n (x_i - \bar{x}), \nu \rangle^2}{m}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$. Often, the labels y lie in a bounded set, at least with high probability. The projected and centered second moments may well be independent of the ambient dimension d, leading to a realistic, even somewhat pessimistic, expectation that

$$L_\lambda(a, W, b) = \frac{1}{n} \sum_{i=1}^{n} \left| f_{(a,W,b)}(x_i) - y_i \right|^2 + \lambda \sum_{i=1}^{m} \left(a_i^2 + |w_i|_{\ell^2}^2 \right) \lesssim \frac{d^2 n^2}{m} + \lambda d n$$

for data sets which do not heavily concentrate at a single point or exhibit heavy tail behavior. While the data can generally be fit exactly if $m > n$ as explained by Llanas and Sainz (2006) this estimate also controls the size of the weights of the neural network needed.

Similarly, this estimate can be used to bound the additional size of the Barron norm which is required to fit values $y_i' = y_i + \varepsilon$, assuming that the Barron norm required to fit y_i is already known. In particular, if the labels are perturbed slightly, it remains possible to fit these by $\tilde{f} = f^* + \sum_{i=1}^n \varepsilon_i \, f_d^*((x - x_i)/r_i)$ where $r_i = \min_{j \neq i} |x_i - x_j|$. The perturbed function $\tilde{f}$ coincides with the true target function f^* except on $\bigcup_{i=1}^n B_{r_i}(x_i)$ and only exceeds its Barron norm by a controlled amount $[f_d^*]_{\mathcal{B}} \sum_{i=1}^n \frac{\varepsilon_i}{r_i}$. Bartlett et al. (2020) and Kornowski et al. (2023) study the situation of benign overfitting where a learned model generalizes well despite fitting noisy labels perfectly see also (Bubeck and Sellke, 2021) for a link between overparametrization and stable interpolation.

5.2 Mollification and Density

Since f_d^* is a compactly supported, non-negative function, it can serve as a mollifier. Namely, for $\varepsilon > 0$ and $u \in L^1_{loc}(\mathbb{R}^d)$, denote

$$\eta_\varepsilon(z) = \frac{f_d^*(z/\varepsilon)}{\|f_d^*\|_{L^1(\mathbb{R}^d)}\varepsilon^d}, \qquad u_\varepsilon(x) = (u * \eta_\varepsilon)(x) = \int_{\mathbb{R}^d} u(z)\,\eta_\varepsilon(x-z)\,\mathrm{d}z.$$

It is well-known that $u_\varepsilon \to u$ in $L^1(K)$ for any compact set $K \subseteq \mathbb{R}^d$ (Dobrowolski, 2010, Lemma 4.22). In many situations, rates can be obtained, either in the L^1-topology or a stronger topology, under the assumption that u lies in a space of more regular functions (e.g. a Hölder or Sobolev space). We consider the following scenario:

Assume that $u \in X$, where $X \subseteq L^1(\mathbb{R}^d)$ is a space of functions $u : \mathbb{R}^d \to \mathbb{R}$ for which it is known that $\|u_\varepsilon - u\|_{L^2(U)} \leq C\|u\|_X \varepsilon^\alpha$ for a given domain $U \subseteq \mathbb{R}^d$ and some universal constants $C, \alpha > 0$ (which may depend on U). If u is naturally defined only on U and not the entire space, extension theorems can often be used to extend u in the same regularity class, see e.g. (Dobrowolski, 2010, Chapter 6).

Note furthermore that u_ε is a continuous superposition of Barron functions in x. Since $\mathcal{B}_0$ is a Banach space, u_ε is a Barron function with norm at most

$$[u_\varepsilon]_{\mathcal{B}} \leq \|u\|_{L^1(\mathbb{R}^d)}\|\eta_\varepsilon\|_{\mathcal{B}} = \frac{\|u\|_{L^1(\mathbb{R}^d)}\,[f_d^*]_{\mathcal{B}}}{\|f_d^*\|_{L^1(\mathbb{R}^d)}\,\varepsilon^{d+1}}.$$

In particular, due to the direct approximation theorem for Barron functions (Proposition 5), there exists a neural network f_m with one hidden layer, ReLU activation, m neurons (and an affine shift) such that

$$\|f_m - u_\varepsilon\|_{L^2(U)} \leq [u_\varepsilon]_{\mathcal{B}}\,\mathrm{meas}(U)\,\mathrm{diam}(U)\,m^{-1/2}$$

and thus

$$\|f_m - u\|_{L^2(U)} \leq \|f_m - u_\varepsilon\|_{L^2(U)} + \|u_\varepsilon - u\|_{L^2(U)} \leq \frac{\|u\|_{L^1(\mathbb{R}^d)}\,[f_d^*]_{\mathcal{B}}\,\mathrm{diam}(U)\,\mathrm{meas}(U)}{\|f_d^*\|_{L^1(\mathbb{R}^d)}\,\varepsilon^{d+1}m^{1/2}} + C\|u\|_X\varepsilon^\alpha$$

Balancing the scaling of terms $\varepsilon^{-(d+1)}m^{-1/2} = \varepsilon^\alpha$, we find that it is optimal to choose $\varepsilon \sim m^{-\frac{1}{2(d+1+\alpha)}}$, which leads to an approximation order of $\varepsilon^\alpha \sim m^{-\frac{\alpha}{2(d+1+\alpha)}}$. We note that not only the rate, but also the constants exhibit the curse of dimensionality. Observe that it is generally impossible to approximate functions in classical function spaces by functions of low norm from any function class in which the unit ball has low Rademacher complexity, so the curse of dimensionality cannot be avoided here (E and Wojtowytsch, 2021).

Since $f_d^* \leq 1$ and $f_d^* \equiv 0$ outside the unit ball, we find that $\|f_d^*\|_{L^1} \leq \omega_d \sim \frac{1}{\sqrt{\pi d}}\left(\frac{2\pi e}{d}\right)^{d/2}$. The true L^1-norm is likely even much smaller, as f_d^* appears to decay rapidly close to the unit sphere. Nevertheless, we find this an easy way to obtain an explicit rate with little effort.

Example 3. *If X is the space of Lipschitz-continuous functions on $\mathbb{R}^d$, then the approximation property holds as*

$$\begin{aligned}\left|u_\varepsilon(x) - u(x)\right| &= \left|\int_{\mathbb{R}^d} \left[u(z) - u(x)\right] \eta(x-z)\,\mathrm{d}z\right| \leq \int_{B_\varepsilon(0)} \eta_\varepsilon(z)|u(x+z) - u(x)|\,\mathrm{d}z \\ &\leq [u]_{Lip} \frac{\int_0^\varepsilon \eta_\varepsilon(r)\, r^d\,\mathrm{d}r}{\int_0^\varepsilon \eta_\varepsilon(r)\, r^{d-1}\,\mathrm{d}r} \leq [u]_{Lip}\varepsilon.\end{aligned}$$

The conditions above are therefore met with $\alpha = 1$.

5.3 Depth Separation

We have seen that any function which satisfies $f \equiv 1$ in $B_\varepsilon(0)$ and $f \equiv 0$ outside of $B_1(0)$ has Barron semi-norm which is exponentially large in the dimension d of the data space (for fixed $\varepsilon \in (0,1)$).

By comparison, the function

$$f(x) = \begin{cases} 1 & |x| \leq \varepsilon \\ 1 - \frac{|x|-\varepsilon}{1-\varepsilon} & \varepsilon \leq |x| \leq 1 \\ 0 & |x| \geq 1 \end{cases}$$

can be represented as the composition $f = f_1 \circ f_2$ of two Barron functions $f_2 : \mathbb{R}^d \to \mathbb{R}$ and $f_1 : \mathbb{R} \to \mathbb{R}$

$$f_2(x) = \frac{|x| - \varepsilon}{1-\varepsilon}, \qquad f_1(z) = \max\left\{\{0, \min\{1-z, 1\}\right\} = \sigma(1-z) - \sigma(-z)$$

with norm

$$[f_2]_{\mathcal{B}} = \frac{1}{1-\varepsilon} \frac{\int_{S^{d-1}} 1\,\mathrm{d}\mathcal{H}^{d-1}}{\int_{S^{d-1}} \sigma(\nu_1)\,\mathrm{d}\mathcal{H}^{d-1}_\nu} \sim \frac{2\sqrt{d}}{1-\varepsilon}, \qquad [f_1]_{\mathcal{B}} = 2.$$

The second norm estimate be easily obtained by Proposition 4, whereas the second can be obtained as in the second step in the proof of 10 in Appendix B—see also (E and Wojtowytsch, 2020a, Section 4).

In particular, by the direct approximation theorem for Barron functions (Proposition 5), it is possible to approximate f_2 with parameters whose magnitude does not exceed $C(1-\varepsilon)d^{-1/2}$. When written as a neural network with two hidden layers, the initial linear layer of f_1 and terminal linear layer of f_2 are concatenated into a single linear map. Balancing the magnitude of coefficients equally over all layers, we find that the parameters scale only like $d^{1/6}(1-\varepsilon)^{-1/3}$, so the weight decay regularizer grows as $d^{1/3}(1-\varepsilon)^{-2/3}$. Observe that weight decay does not induce a norm for deeper ReLU networks due to the mismatch in homogeneities.

Theorem 8 thus serves to illustrate the following depth separation phenomenon: *A function $f : \mathbb{R}^d \to \mathbb{R}$ which takes values 1 on $B_\varepsilon(0)$ and 0 on $\mathbb{R}^d \setminus B_1(0)$ is much easier to approximate by ReLU networks with two hidden layers than with one.* While depth separation phenomena are well established by Eldan and Shamir (2016); Telgarsky (2016); Safran and Shamir (2017), this is a particularly easy criterion. The target function is neither

highly oscillatory nor discontinuous as the data can be fit with Lipschitz-constant $(1-\varepsilon)^{-1}$. The fact that compositions of Barron functions correspond to certain neural networks with two hidden layers has been observed e.g. by E and Wojtowytsch (2020a); Parhi and Nowak (2022). For further observations on the occurrence or non-occurrence of depth separation phenomena, see also Safran et al. (2019); Safran and Lee (2022). In particular, Safran et al. (2019); Venturi et al. (2022) indicate that depth separation depends intricately on how separation is defined and measured.

On the other hand, the result is a weaker version of a depth separations statement than others. We do not claim that the number of neurons required to approximate such a function f to a certain accuracy grows exponentially in dimension, but rather that either the number of neurons or the magnitude of the parameters does. From a practical point of view, both are prohibitive.

6. Finding Optimal Bump Functions

In this section, we compute numerical approximations of the optimal bump functions which were constructed in Theorem 7 for different odd dimensions $d \in \mathbb{N}$ beyond the case $d = 3$ considered in Example 2. As previously, denote $n = \frac{d-1}{2}$, i.e. $d = 2n+1$. For simplicity, we exploit that three tasks are equivalent: Approximating $|s|$ in $L^\infty(-1,1)$ by polynomials of degree at most $2n$ (or $2n+1$), approximating $\sqrt{s}$ in $L^\infty(0,1)$ by polynomials of degree at most n, and approximating s in $L^\infty(0,1)$ by *even* polynomials of degree $2n$. We proceed in three steps:

1. Find the optimal approximation of $s \mapsto \sqrt{s}$ by polynomials of degree n in $L^\infty(0,1)$, and find the $n+2$ points $t_0, \dots, t_{n+1}$ at which the error is maximal. Take the optimal points $s_i = t_i^2$ for the approximation of $f(s) = s$ by *even* polynomials.
2. Solve the linear system (19) to obtain the measure $\mu = \sum_{i=0}^{n+1} \mu_i \delta_{s_i}$. Compute the piecewise linear function g by $g'' = \mu$ in $(0,1]$, $g(0) = 1$ and $g'(0) = \mu_0$.
3. Obtain f from g by numerically integrating (9).

In our implementation, the first step is solved by the Remez algorithm (Trefethen, 2019):

(i) Initialize $s_0, \dots, s_{n+1} \subseteq [0,1]$, e.g. as equi-distant points such that $s_0 = 0$ and $s_1 = 1$.

(ii) Solve the system $\sum_{j=0}^n \alpha_j s_i^j = \sqrt{s} + (-1)^i e$, $0 \leq i \leq n+1$ for the coefficients α_j and the equi-oscillation parameter e.

(iii) Update $s_0, \dots, s_{n+1}$ such that $s_0 = 0$, $s_{n+1} = 1$ and for $i = 1, \dots, n$, s_i is a point at which the unsigned error function $\sqrt{s} - \sum_{j=0}^n \alpha_j s^j$ has a local extremum.

(iv) Iterate (ii) and (iii) until after the final update we have approximately reached equi-oscillating points of largest error:

$$\frac{\max_{0\leq i\leq n+1} \left| \sqrt{s_i} - \sum_{j=0}^n \alpha_j s_i^j \right|}{\min_{0\leq i\leq n+1} \left| \sqrt{s_i} - \sum_{j=0}^n \alpha_j s_i^j \right|} < 1.001.$$

We solve the linear system in step 2 in the iteration scheme by LU factorization. The non-linear system is solved by two nested interval constructions:

- Given $0 = s_0 < \cdots < s_{n+1} = 1$ such that $\sqrt{s_i} - \sum_{j=0}^{n} \alpha_j s_i^j = (-1)^i e$, we conclude that for all $i = 0, \ldots, n$, there exists $t_i \in (s_j, s_{j+1})$ such that

$$\sqrt{t_i} - \sum_{j=0}^{n} \alpha_j t_i^j = 0, \quad i = 0, \ldots, n$$

  by the Intermediate Value Theorem. In particular, the $n+1$ points $0 < t_0 < \cdots < t_n < 1$ are distinct and ordered. We approximate t_i by the bisection method to accuracy $< 10^{-12}$. Note that $e \neq 0$, since the approximating polynomial cannot match the objective function at $n+2$ points by the same argument as in the proof of Theorem 7.

- Given $t_0, \ldots, t_n$, we find that there for $i = 1, \ldots, n$ there exists $\xi_i \in (t_{i-1}, t_i)$ such that

$$\frac{d}{ds}\bigg|_{s=\xi_i} \left(\sqrt{s} - \sum_{j=0}^{n} \alpha_j s^j \right) = 0, \qquad i = 1, \ldots, n$$

  by Rolle's Theorem. Again, we approximate ξ_i by the bisection method to accuracy $< 10^{-12}$. By construction, all ξ_i are distinct. We update

$$\{s_0, s_1, \ldots, s_n, s_{n+1}\} \mapsto \{s_0, \xi_1, \ldots, \xi_n, s_{n+1}\}.$$

The nested interval construction is more numerically stable than Newton-Raphson iteration, as a Newton solver tends to find the same ξ_i multiple times starting at different roots $s_i, s_{i'}$ from the previous iteration.

The linear step (2) is solved by LU factorization. The integral in (3) is evaluated using a composite Simpson rule and 1,001 integration points. A sample implementation of the algorithm can be found in a google colab notebook by Wojtowytsch (2022).

We note that the linear system (19) can be solved for *any* choice of distinct points $0 \leq s_0 < \cdots < s_{n+1} \leq 1$. To explore the importance of using the optimal points found using the Remez algorithm, we compare g and f for the optimal choice of sample points and other, more classic and explicit choices s_i in Figures 1 and 2 respectively. The Barron norm grows slowly and linearly for optimal break points and faster than linearly for other explicit choices of break points, as can be seen in the rightmost image in Figure 3. In the left and middle plot of Figure 3, we also display the known profiles of Barron functions due to Ongie et al. (2019) as well as profiles of Barron functions which are constant in a neighbourhood of the origin.

For any choice of break points which include zero, the function f is a 'wizard's hat' function: Monotone decreasing, flat away from the origin, monotone decreasing and convex, non-smooth at the origin. It is thus qualitatively different from previously known radial profiles due to Ongie et al. (2019).

Additional empirical results relating to the optimal construction can be found in Appendix A.

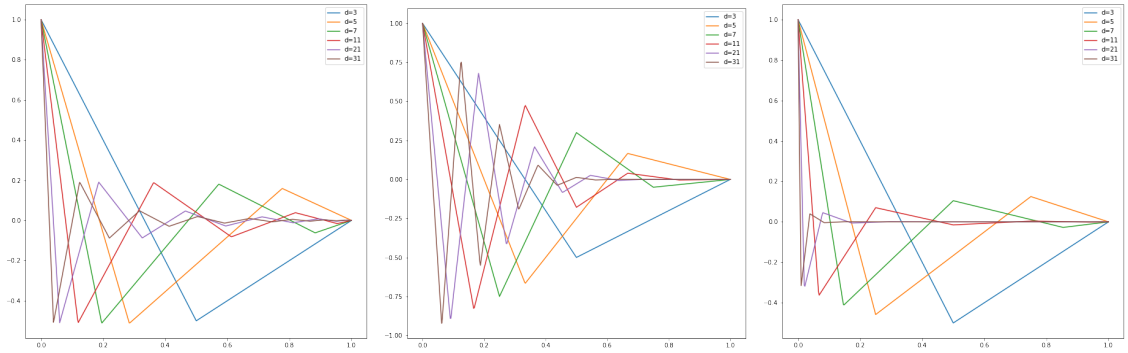

Figure 1: We compute the piecewise linear function g satisfying the moment conditions (19) for three different choices of $n+2 = \frac{d+3}{2}$ break points s_i: Optimal points found by the Remez algorithm (left), equi-distant points $s_i = i/(n+1)$ (middle) and the roots of Chebyshev polynomials of the second kind $s_i = 1/2 + \cos\left(i\pi/(n+1)\right)/2$ (right). For the optimal choice of s_i, we empirically observe that the collection of points $\{(s_i, g_d(s_i)) : d \in 2\mathbb{N}+1\}$ concentrates on a line ℓ_i parallel to the horizontal axis. For equi-distant nodes, the oscillations become larger as d increases, whereas they become smaller for Chebyshev nodes.

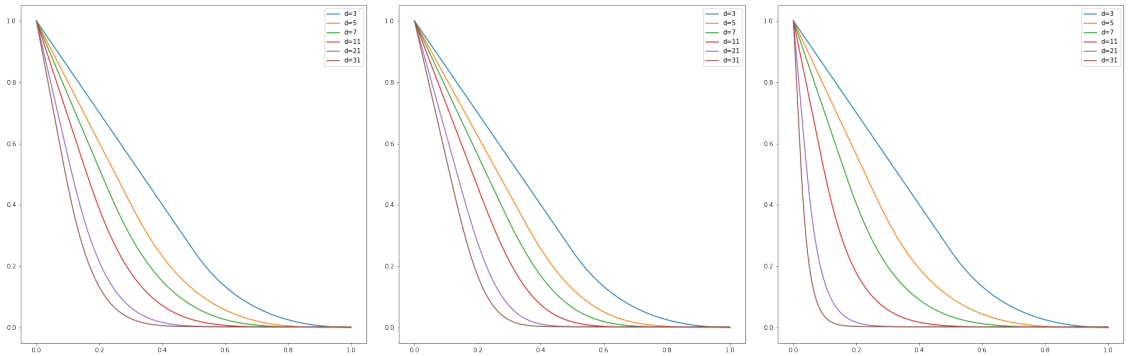

Figure 2: We compare the functions f computed by (9) for the piecewise linear functions g in Figure 1 associated to three different choices of break points s_i: Optimal (left), equi-distant (middle), Chebyshev (right). The break points and curves agree for $d = 3$ (blue curve) and are qualitatively similar for all $d \geq 3$, in particular non-negative, monotone-decreasing and convex. The curves are steeper at the origin in higher dimensions, most noticeably for Chebyshev nodes. The curve with optimal break points appears to make the slowest transition from $f = 1$ to $f \approx 0$.

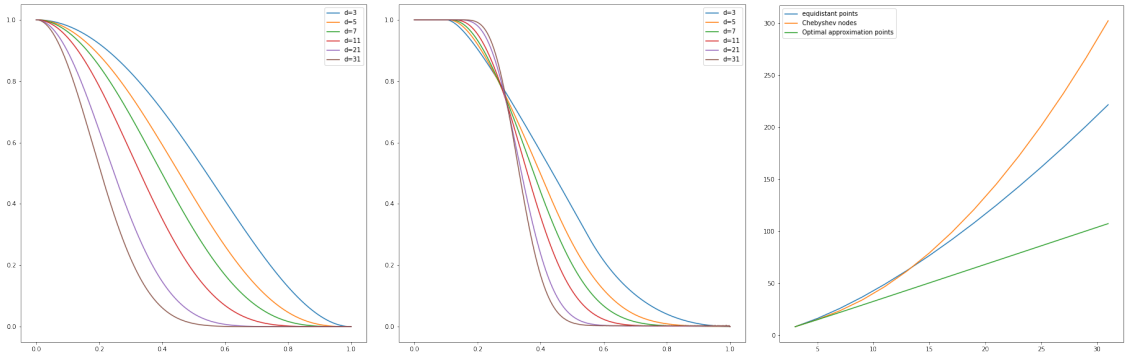

Figure 3: **Left:** The radial profiles of the known Barron bump functions $f(|x|) = (1 - |x|^2)^{\frac{d-1}{2}}$ of Ongie et al. (2019) are smooth at the origin and non-convex in the radial direction. They are thus geometrically distinct from the profiles associated to piecewise linear functions g as depicted in Figure 2. **Middle:** The functions f associated to piecewise linear functions g with break points at equidistant points $s_i = 0.1 + 0.9 \cdot i/(n+1)$ in $[0, 1]$ are C^1-smooth, monotone decreasing and non-negative, but not convex in the radial direction. **Right:** The Barron semi-norm $\sum_{i=0}^{n+1} |\mu_i|$ of functions f associated to g with different break points s_i grows slowest (and linearly) for optimal the optimal choice of points and fastest for break points at Chebyshev nodes. All growth rates are ostensibly polynomial of empirical degree 1.1 (optimal points), 1.4 (equi-distant points) and 1.6 (Chebyshev nodes) as determined by least squares fitting. By comparison, the norm growth for the choice of equi-distant nodes in $[0.1, 1]$ in the middle figure is exponentially large in d and is not pictured for better readability of the plots.

7. Conclusion and Open Problems

We have provided an explicit construction for how neural networks optimally interpolate certain radially symmetric data with respect to a weight decay regularizer in the infinite parameter limit. While we do not prove that the optimal interpolant is radially symmetric, the radial average of all interpolants coincides with the solution constructed in this article. We show that its weight decay regularizer grows as d and its Lipschitz constant grows at most as $\sqrt{d}$. In contrast, we identify a slight modification which necessitates exponential growth. A number of important questions remain open, even for shallow neural networks and the simple case of rotational symmetry. Deeper networks appear to be out of reach for our methodology.

- **Is the radially symmetric minimizer the only one?** A uniqueness statement would allow us to establish that regularized risk minimization does in fact lead to symmetry learning, at least in a toy example, and would allow us to prove stronger convergence results in the companion article (Park et al., 2023). We give further heuristic consideration to this question in Appendix D.

- **What happens if we modify the constraints?** For example, it is not clear from the proof of Theorem 8 whether the constraint $f \geq 1$ on $B_\varepsilon(0)$ induces the curse of dimensionality as the constraint $f \equiv 1$ on $B_\varepsilon(0)$ does. Similarly, it may be interesting to study the case where the boundary condition $f \equiv 0$ is imposed on a shell $\{1 \leq |x| \leq R\}$ rather than the entire exterior domain. We recover the problem studied in this article in the limit $R \to \infty$, whereas the optimal solution in the case $R = 1$ would be $f(x) = 1 - |x|$. Furthermore, a modified minimization problem is required to find optimal mollifiers:

  $$\text{Find } \tilde{f}_d^* \in \operatorname*{argmin}_{f \in \widetilde{\mathcal{F}}} \frac{[f]_{\mathcal{B}}}{\|f\|_{L^1}}, \qquad \widetilde{\mathcal{F}} = \left\{ f \in \mathcal{B}_0(\mathbb{R}^d) \cap C_c(\overline{B_1(0)}) : f \geq 0 \right\}.$$

  It appears that subtle differences may make the difference between a solvable data fitting problem and one where we encounter the curse of dimensionality.

- **What more can we say about the optimal function f_d^*?** For example, we do not provide a lower bound on the Lipschitz constant of f_d^*, nor do we study the decay of $f_d^*(r)$ for fixed $r \in (0,1)$ or $\|f\|_{L^1(\mathbb{R}^d)}$ rigorously. We conjecture that both decay at least exponentially in d, and that both sequences are monotone in d. Limited evidence is provided in Appendix D.

Finally, the fact that the extrema of $g = g_n$ lie on straight lines parallel to the horizontal axis as we vary n appears too specific to be random. It is not clear to us how to interpret this observation.

Acknowledgements

The author would like to thank Jonathan Siegel and Rahul Parhi for inspiring conversations.

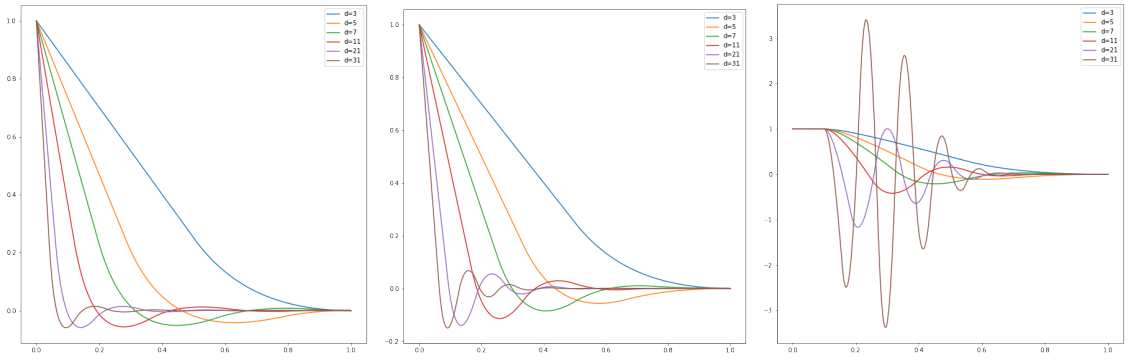

Figure 4: We plot the radial profile of $f_{1,n} : \mathbb{R}^3 \to \mathbb{R}$ as in (12) for various choices of $n = \frac{d-1}{2}$ break points. The points are chosen optimally (for dimension $d = 2n+1$) on the left, equi-distant in $[0, 1]$ in the middle plot and equidistant in $[0.1, 1]$ on the right. Notably, the functions are neither monotone nor non-negative if $d > 3$, and the number of local extrema increases as n grows.

Appendix A. Further Plots

We note the following: If g is a piecewise linear function with n break points which satisfies the $n + 2$ linear moment conditions (16), then it also satisfies the same linear moment conditions for any $m \leq n$. In particular, the function

$$f_{m,n} : \mathbb{R}^{2m+1} \to \mathbb{R}, \qquad f_{m,n}(x) = \frac{\int_{-1}^{1} g\big(|x|s\big)\,(1-s^2)^{\frac{m-3}{2}}\,\mathrm{d}s}{\int_{-1}^{1}(1-s^2)^{\frac{m-3}{2}}\,\mathrm{d}s} \tag{12}$$

is a Barron function such that $f_{m,n}(0) = 1$ and $f_{m,n} \equiv 0$ on $\mathbb{R}^k \setminus B_1(0)$ for $k \leq n$. We plot $f_{m,n}$ for $m = 1$ (i.e. $2m+1 = 3$) and various choices of n and various choices of break points in Figure 4. In Figure 5 we fix $n = 10$ instead and consider the influence of varying m.

The larger the discrepancy between m and n, the more oscillatory the function $f_{m,n}$ is. This is reminiscent of observations from Step 4 in the proof of Theorem 7.

In Figure 6, we numerically investigate the decay of f_d^* as d varies, both pointwise in r and integrated. Since f_d^* is $\sqrt{d}$-Lipschitz and $f_d^*(0) = 1$, we see that the one-dimensional integral $\int_0^1 f_d^*(r)\,\mathrm{d}r$ is bounded from below by $\Omega(d^{-1/2})$. This indeed appears to be the dominant term, and for fixed $r > 0$, we observe empirically that $f_d^*(r)$ decays to zero exponentially in d.

Appendix B. Postponed Proofs

Recall the co-area formula, which allows us to integrate over a Riemannian manifold M by 'slicing' the domain into the level sets of a function $\phi : M \to N$, where N is another Riemannian manifold (Burago and Zalgaller, 2013, Theorem 13.4.2). In the case of slicing

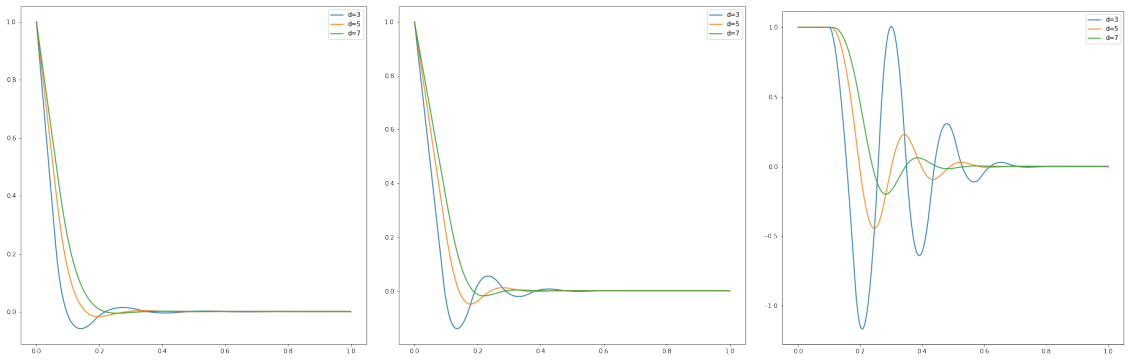

Figure 5: We plot $f_{m,n}$ as in (12) corresponding to low dimension $d \in \{3, 5, 7\}$, $m = \frac{d-1}{2}$ and $n = 10$. and various choices of break points: Optimal for $n = 10$ (left), equidistant in $[0, 1]$ (middle) and equi-distant in $[0.1, 1]$ (right). The radial profiles of the Barron functions are neither monotone nor non-negative. The number of local extrema of the profiles is larger if the dimension d is small compared to the number of break points. The oscillations are smallest for the optimal choice of break points and largest for break points which are bounded away from the origin.

the sphere into level sets of a coordinate projection $\phi(x) = x_1$, the formula reads as

$$\int_{S^{d-1}} f(x) \, \mathrm{d}\mathcal{H}_x^{d-1} = \int_{-1}^{1} \left(\int_{S^{d-1} \cap \{x_1 = s\}} f(x) \, \mathrm{d}\mathcal{H}_x^{d-2} \right) (1 - s^2)^{-1/2} \, \mathrm{d}s$$

since $1 - x_1^2 = |\nabla^{\|} \phi|^2$ is the modulus of the tangential gradient of ϕ, which measures volume distortions. This can be considered a curvinlinear version of Fubini's theorem. If f only depends on x_1, the formula further simplifies to

$$\int_{S^{d-1}} f(x) \, \mathrm{d}\mathcal{H}_x^{d-1} = (d-1)\omega_{d-1} \int_{-1}^{1} f(s, 0, \dots, 0) \, (1 - s^2)^{(d-3)/2} \, \mathrm{d}s \tag{13}$$

since $S^{d-1} \cap \{x_1 = s\}$ is a $d-2$-dimensional Euclidean sphere of radius $\sqrt{1 - s^2}$. Here ω_{d-1} denotes the volume of the $d-1$-dimensional unit ball and $(d-1)\omega_{d-1}$ the volume of the $d-2$-dimensional Euclidean unit sphere.

The first proof we give in this Section is for Lemma 10.

Proof Step 1. Symmetrization. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a radially symmetric Barron function. Then in particular $f(x)$ is the same as the average over $f(Ox)$ for O in $SO(d)$ and the average is taken with respect to the Haar measure H (which coincides with the

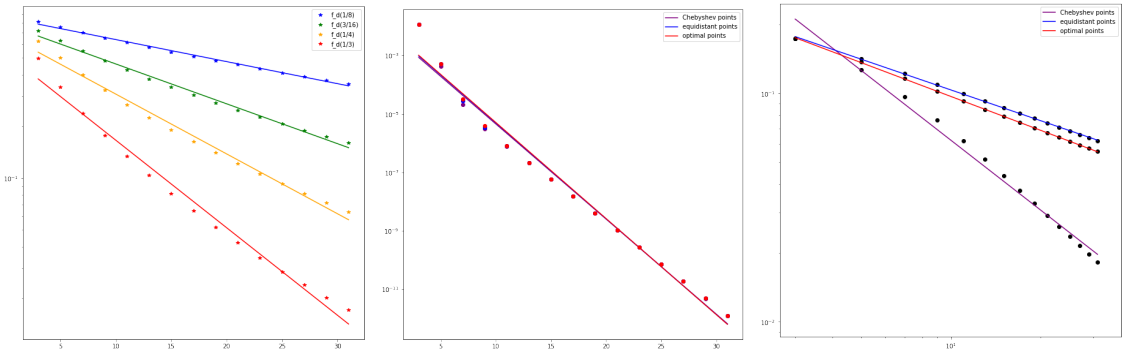

Figure 6: **Left:** We plot $f_d^*(x)$ as a function of d for fixed x in a logarithmic scale together with $\exp(\alpha(x)d+\beta(x))$, where α, β are chosen depending on x as the least squares fit for the function $\log(f_d^*(x))$. The graphs suggest that the decay is exponential in d and thus comparable to explicit solutions $\tilde{f}_d(x) = (1-|x|^2)^{\frac{d+3}{2}}$ of Ongie et al. (2019). **Middle:** We graphically compare the decay of the normalized d-dimensional integral of $⨍_{B_1(0)} f_d(x)\,\mathrm{d}x$ for different choices of break points. Despite graphical differences around the origin, the values of the integrals are very similar and decay roughly as $\exp(-4.7 - 0.75 * d)$. In dimension three, all three functions f_d coincide, while their difference close to the origin becomes negligible in high dimension, where almost all measure concentrates by the boundary of the unit ball. **Right:** We graphically compare the decay of the 1-dimensional integral $\int_0^1 f_d(r \cdot e_1)\,\mathrm{d}r$ of the function f_d associated to piecewise linear g with $n = \frac{d-1}{2}$ break points in $(0,1)$ for different choices of break points. The integral empirically decays as $0.3 \cdot d^{-0.49}$ for optimal points, like $0.29 \cdot d^{-0.45}$ for equidistant points and like $0.64 \cdot d^{-1.02}$ for Chebyshev points. The order of decay was established by a least squares regression.

$\frac{d(d-1)}{2}$-dimensional Hausdorff measure on $SO(d)$ with respect to the Frobenius norm), i.e.

$$\begin{aligned} f(x) &= \fint_{SO(d)} f(Ox)\,\mathrm{d}H_O \\ &= \fint_{SO(d)} \int_{\mathbb{R}^{d+2}} \sigma(w^T Ox + b)\,\mathrm{d}\mu_{(w,b)}\,\mathrm{d}H_O \\ &= \int_{\mathbb{R}^{d+2}} \fint_{SO(d)} \sigma\big((O^T w)^T x + b\big)\,\mathrm{d}H_O\,\mathrm{d}\mu_{(w,b)} \\ &= \int_{\mathbb{R}^{d+2}} |w| \fint_{S^{d-1}} \sigma\left(\nu^T x + \frac{b}{|w|}\right)\mathrm{d}\mathcal{H}^{d-1}_\nu\,\mathrm{d}\mu_{(w,b)} \end{aligned}$$

since for any $w \in S^{d-1}$, the map $SO(d) \to S^{d-1}$, $O \mapsto Ow$ pushes the Haar measure forward to the uniform distribution on S^{d-1}. Thus f can be written as a continuous linear combination of the elementary radially symmetric Barron functions

$$f_b(x) = \fint_{S^{d-1}} \sigma\big(\nu^T x - b\big)\,\mathrm{d}\mathcal{H}^{d-1}_\nu \quad \text{and } f_\infty(x) \equiv 1.$$

On the other hand, every function of this type is a radially symmetric Barron function. Finally, we note that

$$\begin{aligned} f_b(x) - f_{-b}(x) &= \fint_{S^{d-1}} \sigma\big(\nu^T x - b\big) - \sigma\big(\nu^T x + b\big)\,\mathrm{d}\mathcal{H}^{d-1}_\nu \\ &= \fint_{S^{d-1}} \sigma\big(-\nu^T x - b\big) - \sigma\big(\nu^T x + b\big)\,\mathrm{d}\mathcal{H}^{d-1}_\nu \\ &= \fint_{S^{d-1}} \nu^T x + b\,\mathrm{d}\mathcal{H}^{d-1}_\nu \\ &= b, \end{aligned}$$

since the uniform distribution is invariant under the substitution $\nu \mapsto -\nu$ in the first term. In particular, every radially symmetric Barron function can be written as

$$f(x) = f(0) + \int_{[0,\infty)} f_b(x)\,\mathrm{d}\mu_b \tag{14}$$

for some measure μ on $[0,\infty)$ and $[f]_{\mathcal{B}} = \|\mu\|_{TV}$, since $f_b(0) = 0$ for any $b > 0$.

Step 2. Gradient bound. We note that $[f_b]_{\mathcal{B}} = 1$ for any b by definition and

$$\nabla f_b(x) = \fint_{S^{d-1}} \sigma'\big(\nu^T x - b\big)\,\nu\,\mathrm{d}\mathcal{H}^{d-1}_\nu = \fint_{S^{d-1}} 1_{\{\nu^T x > b\}}\nu\,\mathrm{d}\mathcal{H}^{d-1}_\nu.$$

Due to radial symmetry, the gradient points in direction x, i.e.

$$\nabla f_b(x) = \frac{\int_{S^{d-1}} 1_{\{\nu_1 > b/|x|\}}\nu_1\,\mathrm{d}\mathcal{H}^{d-1}_\nu}{\int_{S^{d-1}} 1\,\mathrm{d}\mathcal{H}^{d-1}_\nu}\,\frac{x}{|x|}$$

The gradient is largest as $|x| \to \infty$, and

$$\sup_{x\in\mathbb{R}^d} |\nabla f_b(x)| = \frac{\int_{S^{d-1}} 1_{\{\nu_1>0\}}\nu_1\,\mathrm{d}\mathcal{H}^{d-1}_\nu}{\int_{S^{d-1}} 1\,\mathrm{d}\mathcal{H}^{d-1}_\nu} = \frac{\int_{S^{d-1}} 1_{\{\nu_1>0\}}\nu_1\,\mathrm{d}\mathcal{H}^{d-1}_\nu}{2\int_{S^{d-1}} 1_{\{\nu_1>0\}}\,\mathrm{d}\mathcal{H}^{d-1}_\nu} = \frac{\int_0^1 s(1-s^2)^{\frac{d-3}{2}}\,\mathrm{d}s}{2\int_0^1 (1-s^2)^{\frac{d-3}{2}}\,\mathrm{d}s}$$

independently of b. In the last step, we used the coarea formula (13). It is now possible to evaluate the gradient

$$\frac{\int_0^1 s(1-s^2)^{\frac{d-3}{2}}\,\mathrm{d}s}{2\int_0^1(1-s^2)^{\frac{d-3}{2}}\,\mathrm{d}s} = \frac{\frac{1}{d-1}}{\sqrt{\pi}\frac{\Gamma\left((d-1)/2\right)}{\Gamma(d/2)}} = \frac{\Gamma(d/2)}{\sqrt{\pi}\,(d-1)\,\Gamma\left((d-1)/2\right)} \sim \frac{1}{\sqrt{2\pi d}}$$

in the sense that

$$\lim_{d\to\infty}\sqrt{d}\frac{\int_0^1 s(1-s^2)^{\frac{d-3}{2}}\,\mathrm{d}s}{2\int_0^1(1-s^2)^{\frac{d-3}{2}}\,\mathrm{d}s} = \frac{1}{\sqrt{2\pi}}.$$

Consequently, for a general radially symmetric Barron function as in (14) and sufficiently large $d \in \mathbb{N}$, we find that

$$[f]_{Lip} = \sup_{x\in\mathbb{R}^d}\left|\nabla f(x)\right| \leq \int_{[0,\infty)} \|\nabla f_b\|_{L^\infty}\,\mathrm{d}|\mu|_b \leq \frac{1+\varepsilon}{\sqrt{2\pi d}}\|\mu\|_{TV} = \frac{1+\varepsilon}{\sqrt{2\pi d}}[f]_{\mathcal{B}}.$$

Step 3. Higher regularity. By Corollary 3, any radially symmetric Barron function is C^1-smooth except at the origin. This establishes the claim in the case $d = 3$.

Note that if $f : (0,\infty) \to \mathbb{R}$ is C^k-smooth, then the same is true for $F : \mathbb{R}^d \to \mathbb{R}$ given by $F(x) = f(|x|)$ by the chain rule and product rule. It thus suffices to analyze the radial profile f of F. In the following, we will denote both functions as f by a slight abuse of notation. Consider the radial profile of the function

$$f_b(r) = c_d\int_0^1 \sigma(sr-b)\,(1-s^2)^{\frac{d-3}{2}}\,\mathrm{d}s, \qquad c_d = \frac{1}{\int_{-1}^1(1-s^2)^{\frac{d-3}{2}}\,\mathrm{d}s}$$

for $b > 0$. We can compute the first two derivatives of f_b by exchanging differentiation and integration

$$\begin{aligned} f_b'(r) &= c_d\int_0^1 \sigma'(sr-b)\,s(1-s^2)^{\frac{d-3}{2}}\,\mathrm{d}s \\ f_b''(r) &= c_d\int_0^1 \sigma''(sr-b)\,s^2(1-s^2)^{\frac{d-3}{2}}\,\mathrm{d}s \quad = \frac{c_d}{r}\big(b/r\big)^2\max\left\{1-\big(b/r\big)^2,0\right\}^{\frac{d-3}{2}}, \end{aligned}$$

where the second formula must be justified by approximation, as the derivative $\frac{d^2}{dr^2}\sigma(sr-b) = \frac{1}{r}\cdot\delta_{b/r}$ (considered as a 'function' of s) is not regular. For $b > 0$, it is easy to see that $f_b \equiv 0$ is C^∞-smooth in $[0,b)$, and as a polynomial in $1/r$ also C^∞-smooth on (b,∞). Clearly f_b'' and all its derivatives vanish at infinity. If $d = 3$, f_b'' is continuous except at $r = b$, where it has a jump discontinuity. If $d \geq 5$, the function

$$f_b''(r) = c_d\, r^{-d}b^2(r^2-b^2)^{\frac{d-3}{2}} = c_d b^2\, r^{-d}(r-b)^{\frac{d-3}{2}}(r+b)^{\frac{d-3}{2}}$$

vanishes as $(r-b)^{\frac{d-3}{2}}$ at $r = b$ and thus has $\frac{d-5}{2}$ additional derivatives which vanish at $r = b$. We find $f_b \in C^{\frac{d-1}{2}}$ for any odd dimension d. The $\frac{d+1}{2}$-th derivative of f_b is bounded and continuous except at $r = b$, and the $\frac{d+3}{2}$-th derivative of f_b'' is a finite measure associated to the regular part of the derivative in (b,∞) and the jump at $r = b$.

It remains to show that a general radial Barron function

$$f(r) = f(0) + \int_{[0,\infty)} f_b(r)\,\mathrm{d}\mu_b$$

has the same regularity as its components f_b, at least away from the origin. To simplify the presentation, we focus on the case $d \geq 5$. Let $\varepsilon > 0$ and observe that

$$f(r) = f(0) + \mu(\{0\})\, f_0(r) + \int_{(0,\varepsilon]} f_b(r)\,\mathrm{d}\mu_b + \int_{(\varepsilon,\infty)} f_b(r)\,\mathrm{d}\mu_b.$$

Clearly, the affine linear component $f(0) + \mu(\{0\})\, f_0(r)$ is C^∞-smooth except at the origin. Secondly, we note that for any $b > 0$, the identity $f_b(r) = b\, f_1\big(r/b\big)$ holds. In particular,

$$\frac{d^k}{dr^k} \int_{(\varepsilon,\infty)} f_b(r)\,\mathrm{d}\mu_b = \int_{(\varepsilon,\infty)} b^{1-k}\, f_1^{(k)}\left(\frac{r}{b}\right)\,\mathrm{d}\mu_b$$

where the integrals converge uniformly for $k \leq \frac{d-1}{2}$ due to the L^∞-bound on the $k+1$-th derivative of f_b. Similarly, the $\frac{d+1}{2}$-th derivative converges in L^p for all $p < \infty$ due to the bound on the measure-valued $\frac{d+3}{2}$-th derivative. Finally, for $k = \frac{d+3}{2}$, the integral converges weakly in the sense of Radon measures, i.e. in the weak-* sense, when we consider the space of (Radon) measures as dual to the space of continuous functions.

For the first integral, we prove convergence assuming that $r \geq \varepsilon$. Note that $f_b''(r) = r^{-1}\, P(b/r)$ for some polynomial P and $r \geq \varepsilon \geq b$. By induction we see that $f_b^{(k)}(r) = r^{1-k} P_k(b/r)$ for all $k \geq 2$, where P_k is another polynomial. This is easily seen since

$$\begin{aligned}\frac{d}{dr}\big[r^{1-k} P_k(b/r)\big] &= (1-k)\, r^{-k}\, P_k(b/r) + r^{1-k}\, P_k'(b/r)\, \left(-\frac{b}{r^2}\right) \\ &= r^{-k}\left((1-k)\, P_k(b/r) - \frac{b}{r}\, P_k'(b/r)\right).\end{aligned}$$

Hence, as before,

$$\frac{d^k}{dr^k} \int_{(0,\varepsilon]} f_b(r)\,\mathrm{d}\mu_b = \int_{(0,\varepsilon]} r^{1-k} P_k(b/r)\,\mathrm{d}\mu_b = r^{1-k} \int_{(0,\varepsilon]} P_k(b/r)\,\mathrm{d}\mu_b.$$

The integral converges since $b/r \in [0,1]$ and P_k is a continuous function. ∎

We now come to the proof of Lemma 11.

Proof First claim. Assume that $g : \mathbb{R} \to \mathbb{R}$ is a function with the properties outlined above and f is defined by (9). Then f is radially symmetric by definition and

$$f(0) = \frac{\int_{-1}^1 (1-s^2)^{\frac{d-3}{2}}\, g(0)\,\mathrm{d}s}{\int_{-1}^1 (1-s^2)^{\frac{d-3}{2}}\,\mathrm{d}s} = 1.$$

Furthermore, if $r = |x| \geq 1$ and $\tilde{c}_d := \int_{-1}^{1} (1-s^2)^{\frac{d-3}{2}} \, \mathrm{d}s$, then

$$\begin{aligned}
\tilde{c}_d \, f(x) &= \int_{-1}^{1} (1-s^2)^{\frac{d-3}{2}} \, g(rs) \, \mathrm{d}s \\
&= r^{2-d} \int_{-1}^{1} (r^2 - (rs)^2)^{\frac{d-3}{2}} \, g(rs) \, r \, \mathrm{d}s \\
&= r^{2-d} \int_{-r}^{r} (r^2 - z^2)^{\frac{d-3}{2}} \, g(z) \, \mathrm{d}z \\
&= r^{2-d} \int_{-1}^{1} (r^2 - z^2)^{\frac{d-3}{2}} \, g(z) \, \mathrm{d}z
\end{aligned}$$

since $g \equiv 0$ outside of $(-1, 1)$. The integral vanishes by (8) if $d \geq 3$ is odd, since $(r^2 - z^2)^{\frac{d-3}{2}}$ is an even polynomial of degree at most $d-3$ for all r.

It remains to show that f is a Barron function. Take any measure μ such that

$$g(x) = \int_{-\infty}^{\infty} \sigma(x+b) \, \mathrm{d}\mu_b$$

as in Proposition 4 and compute that

$$\begin{aligned}
\int_{-\infty}^{\infty} \fint_{S^{d-1}} \sigma(\nu^T x + b) \, \mathrm{d}\mathcal{H}^{d-1} \, \mathrm{d}\mu_b &= \fint_{S^{d-1}} \int_{-\infty}^{\infty} \sigma(\nu^T x + b) \, \mathrm{d}\mu_b \, \mathrm{d}\mathcal{H}^{d-1}_{\nu} \\
&= \fint_{S^{d-1}} g(\nu^T x) \, \mathrm{d}\mathcal{H}^{d-1}_{\nu} \\
&= \frac{|S^{d-2}|}{|S^{d-1}|} \int_{-1}^{1} g(s) \, (1-s^2)^{\frac{d-3}{2}} \, \mathrm{d}s \quad = f(x)
\end{aligned} \tag{15}$$

by the co-area formula (13). The fact that the normalizing constant is exactly

$$\tilde{c}_d = \frac{|S^{d-2}|}{|S^{d-1}|} = \int_{-1}^{1} (1-s^2)^{\frac{d-3}{2}} \, \mathrm{d}s$$

can be justified by the same co-area integration. Finally, we note that the left hand side of (15) is clearly a radially symmetric Barron function satisfying $[f]_{\mathcal{B}} \leq \|\mu\|_{TV}$. Taking the infimum over all μ representing g, we find that $[f]_{\mathcal{B}(\mathbb{R}^d)} \leq [g]_{\mathcal{B}(\mathbb{R})}$.

Second claim. Assume on the other hand that f is a radially symmetric Barron function. If we denote by $\overline{\mu}$ the Haar measure on the special orthogonal group $SO(d)$, then

due to radial symmetry

$$\begin{aligned} f(x) &= \int_{SO(d)} f(Ox)\,\mathrm{d}\overline{\mu}_O \\ &= f(0) + \int_{SO(d)} \int_{\mathbb{R}^{d+1}} \sigma(w^T Ox + b)\,\mathrm{d}\mu_{(w,b)}\,\mathrm{d}\overline{\mu}_O \\ &= f(0) + \int_{\mathbb{R}^{d+1}} \fint_{SO(d)} \sigma\big((O^T w)^T x + b\big)\,\mathrm{d}\overline{\mu}_O\,\mathrm{d}\mu_{(w,b)} \\ &= f(0) + \int_{\mathbb{R}^{d+1}} |w| \fint_{S^{d-1}} \sigma\left(\nu^T x + \frac{b}{|w|}\right)\mathrm{d}\mathcal{H}^{d-1}_\nu \mathrm{d}\mu_{(w,b)} \\ &= f(0) + \int_{-\infty}^{\infty} \fint_{S^{d-1}} \sigma\left(\nu^T x + b'\right)\,\mathrm{d}\mathcal{H}^{d-1}_\nu\,\mathrm{d}\hat{\mu}_{b'} \end{aligned}$$

where $\hat{\mu} = \Phi_\sharp(|w| \cdot \mu)$ for the map

$$\Phi : \mathbb{R}^{d+1} \to \mathbb{R}, \qquad \Phi(w,b) = \frac{b}{|w|}.$$

It is now possible to reverse the calculations from Step 1 by setting

$$g : \mathbb{R} \to \mathbb{R}, \qquad g(x) = \int_{\mathbb{R}} \sigma(x-b)\,\mathrm{d}\hat{\mu}_b.$$

Taking the infimum over μ representing f, we find that $[g]_{\mathcal{B}(\mathbb{R})} \leq [f]_{\mathcal{B}(\mathbb{R}^d)}$. Clearly, both $s \mapsto g(s)$ and $s \mapsto g(-s)$ induce the same function f by (9) due to symmetry, and so does the even representative $s \mapsto \big(g(s) + g(-s)\big)/2$.

It remains to show that $g \equiv 0$ outside of $(-1,1)$ and that the moment conditions (16) hold. Assuming that $g \equiv 0$ outside $(-1,1)$, we find that

$$\begin{aligned} 0 &= \int_{-1}^1 (1-s^2)^{\frac{d-3}{2}} g(rs)\,\mathrm{d}s = r^{2-d} \int_{-1}^1 \big(r^2 - (rs)^2\big)^{\frac{d-3}{2}} g(rs)\, r\,\mathrm{d}s \\ &= r^{2-d} \int_{-1}^1 (r^2 - z^2)^{\frac{d-3}{2}}\, g(z)\,\mathrm{d}z \end{aligned}$$

for all $r \geq 1$, as the integral over $(-r,-1) \cup (1,r)$ vanishes. The moment conditions follow easily as

$$\frac{d^k}{dr^k} \int_{-1}^1 (r^2 - z^2)^{\frac{d-3}{2}}\, g(z)\,\mathrm{d}z = \frac{d^k}{dr^k} \sum_{j=0}^{(d-3)/2} \binom{(d-3)/2}{j} r^{d-3-2j} \int_{-1}^1 g(z)\, z^{2j}\,\mathrm{d}z \equiv 0$$

for $r \in [1,\infty)$ and $k \geq 1$. Taking $k = \frac{d-3}{2}$ derivatives, we find that g is L^2-orthogonal to z^0. Lowering the order of the derivative inductively, we find that g is L^2-orthogonal to all even polynomials of degree at most $d-3$.

Thus we only need to show that $g \equiv 0$ outside $(-1,1)$. First consider the case $d = 3$, i.e. $n = 0$ and thus

$$f(r) = \int_{-1}^1 g\big(rs\big)\,\mathrm{d}s.$$

Then for $r \geq 1$, we have

$$0 = f(r) = \frac{1}{r}\int_{-1}^{1} g\big(rs\big)\, r\,\mathrm{d}s = \frac{1}{r}\int_{-1}^{1} g(z)\,\mathrm{d}z + \frac{1}{r}\int_{1}^{r} g(z)+g(-z)\,\mathrm{d}z = \frac{1}{r}\int_{1}^{r} g(z)+g(-z)\,\mathrm{d}z$$

since $\int_{-1}^{1} g(s)\,\mathrm{d}s = f(1) = 0$. As g is an even function, we conclude that $\int_{1}^{r} g(s)\,\mathrm{d}s = 0$ for all $r \geq 1$ and thus $g(s) = 0$ for all $s > 1$. We now proceed inductively: Assume that $n \geq 1$ is such that

$$r^{-(n+1)}\int_{-r}^{r} g(z)(r^2-z^2)^n\,\mathrm{d}z = r^{-(n+1)}\int_{-1}^{1} g(rs)(r^2-(rs)^2)^n r\,\mathrm{d}s = \int_{-1}^{1} g(rs)(1-s^2)^n\,\mathrm{d}s = 0$$

for all $r \geq 1$. Then also

$$0 \equiv \int_{-r}^{r} g(z)(r^2-z^2)^n\,\mathrm{d}z \quad \Rightarrow \quad 0 \equiv \frac{d}{dr}\int_{-r}^{r} g(z)(r^2-z^2)^n\,\mathrm{d}z = 2r\int_{-r}^{r} g(z)(r^2-z^2)^{n-1}\,\mathrm{d}z$$

since the boundary term vanishes for $n \geq 1$. In particular, we conclude that

$$\int_{-1}^{1} g(rs)(1-s^2)^n\,\mathrm{d}s \equiv 0 \quad \Rightarrow \quad \int_{-1}^{1} g(rs)(1-s^2)^{n-1}\,\mathrm{d}s \equiv 0$$

for $r \geq 1$ and $n \geq 1$. Since d is odd, we can reduce the integer exponent $n = \frac{d-3}{2}$ inductively until $n = 0$. Then, by the same consideration as in the case $d = 3$, the result is proved. ■

We now prove the abstract statement about measures on the unit interval given in Lemma 12.

Proof Lower bound. Let μ be a finite signed measure satisfying the moment conditions

$$\int_{0}^{1} s\,\mathrm{d}\mu_s = 1, \qquad \int_{0}^{1} s^{2k}\,\mathrm{d}\mu_s = 0 \quad \forall\ 0 \leq k \leq n-1 \tag{16}$$

Then

$$\int_{0}^{1} s\,\mathrm{d}\mu_s = \int_{0}^{1}\left(s - \sum_{k=0}^{n} a_k s^{2k}\right)\mathrm{d}\mu_s \leq \left\| s - \sum_{k=0}^{n} a_k s^{2k}\right\|_{L^\infty(0,1)} |\mu|([0,1]) \tag{17}$$

by definition. Taking the infimum over the parameters $a_0, \ldots, a_n$ on the right, we find that

$$1 = \int_{-1}^{1} s\,\mathrm{d}\mu_s \leq \mathrm{dist}_{L^\infty(0,1)}\big(s \mapsto s,\ \mathrm{span}\{1, s^2, \ldots, s^{2n}\}\big)\cdot\|\mu\|$$

i.e.

$$\|\mu\| \geq \frac{1}{\mathrm{dist}_{L^\infty(0,1)}\big(s \mapsto s,\ \mathrm{span}\{1, s^2, \ldots, s^{2m}\}\big)} = \frac{1}{\mathrm{dist}_{L^\infty(-1,1)}\big(s \mapsto |s|,\ \mathrm{span}\{1, s, \ldots, s^{2m}\}\big)}.$$

The asymptotics of

$$\beta_n := \mathrm{dist}_{L^\infty(-1,1)}\big(s \mapsto |s|, \mathrm{span}\{1, s, \ldots, s^{2m}\}\big)$$

are known due to Bernstein (1912) and Varga and Carpenter (1985) who proved that $\lim_{n\to\infty} n\,\beta_n =: \beta \approx 0.28$, so

$$\liminf_{n\to\infty} \frac{\gamma_n}{n} \geq \liminf_{n\to\infty} \frac{1}{n\beta_n} = \frac{1}{\beta} \approx 3.57.$$

Upper bound: Step 0. Note that due to compactness, there exist parameters $a_0, \dots, a_n$ such that

$$\left\| s - \sum_{k=0}^{n} a_k s^{2k} \right\|_{L^\infty(0,1)} = \mathrm{dist}_{L^\infty(0,1)}\big(s \mapsto s,\ \mathrm{span}\{1, s^2, \dots, s^{2m}\}\big).$$

We fix $a_0, \dots, a_n$ accordingly. Further note that equality is attained in (17) if the measure μ is supported on the set of points

$$\Theta := \left\{ s \in [0,1] : \left| s - \sum_{k=0}^{n} a_k s^{2k} \right| = \max_{r\in[0,1]} \left| r - \sum_{k=0}^{n} a_k r^{2k} \right| \right\}$$

and the measure

$$\widetilde{\mu} = \left(s - \sum_{k=0}^{n} a_k s^{2k} \right) \cdot \mu \tag{18}$$

which has density $s - \sum_{k=0}^{n} a_k s^{2k}$ with respect to μ is non-negative, i.e. μ has "the right sign" at all points. If such a μ exists, it therefore serves as a matching upper bound and the Lemma is proved. It is, however, not immediately clear whether there exists a signed measure μ supported on Θ which satisfies the moment conditions (16) and positivity condition (18). In the following, we will prove that μ indeed does exist.

Step 1. Due to compactness, Θ is a non-empty subset of $[0,1]$. Additionally

$$\Theta \subseteq \{0,1\} \cup \left\{ s \in \mathbb{R} : 2\sum_{k=1}^{n} k a_k\, s^{2k-1} = 1 \right\}$$

since the function $s \mapsto s - \sum_{k=0}^{n} a_k s^{2k}$ is either maximal or minimal at $s \in \Theta$. By the fundamental theorem of algebra, $\Theta = \{s_1, \dots, s_N\}$ is thus a finite subset of $[0,1]$. In this step, we prove that $0, 1 \in \Theta$ and $\Theta \cap (0,1) = n$.

Note that $\sum_{k=1}^{n} a_k s^{2k}$ is also an optimal polynomial approximation of the function $h(s) = |s|$ in $C^0[-1,1]$ in the space $\mathcal{P}_{2n+1}$ of polynomials of degree at most $2n+1$, since the optimal approximation is an even polynomial. By Chebyshev's equi-oscillation Theorem (Kincaid et al., 2009, Section 6.9), there exist $N \geq 2n+3$ distinct points $t_1 < \dots < t_N$ such that the error

$$e(s) = |s| - \sum_{k=0}^{n} a_k s^{2k}$$

satisfies

$$|e(t_i)| = \max_{s\in[-1,1]} |e(s)| \quad \forall\ i = 1, \dots, N \qquad \text{and } e(t_i)e(t_{i+1}) < 0 \quad \forall\ i = 1, \dots, N-1,$$

i.e. there are $N \geq 2n+3$ distinct points where the deviation from the target function is largest, and the oscillation around the target function at consecutive points t_i, t_{i+1} goes in opposite directions.

Clearly, if e is maximal at $s \in [-1,0]$ if and only if it is maximal at $(-s) \in [0,1]$. Therefore, there exist at least $\lceil N/2 \rceil = \lceil (2n+3)/2 \rceil = n+2$ points in $\Theta = [0,1] \cap \operatorname{argmax} e$. Rounding up is required since $2n+3$ is odd, and the point 0 counts fully towards $\Theta \subset [0,1]$. Thus $|\Theta| \geq n+2$.

It remains to show that $|\Theta| \leq n+2$. We prove this only if $n \geq 1$, as the case $n = 0$ of approximation by constant functions can be solved explicitly by direct inspection by the constant polynomial $a_0 = 1/2$.

Assume for a contradiction that $|\Theta| \geq n+3$. Then there exist at least $n+1$ distinct points in $\Theta \cap (0,1)$. Since e is either maximal at $s \in \Theta \cap (0,1)$, we conclude that $e'(s) = 0$ for every $s \in \Theta \cap (0,1)$. By Rolle's Theorem, between any two points s, s' such that $e'(s) = e'(s')$, there exists $s^* \in (s, s')$ such that $e''(s^*) = 0$. In particular, e'' has at least n distinct zeros in $(0,1)$. Since e'' is even, it follows that e'' has at least $2n$ distinct zeros. But, since e'' is a polynomial of degree $2n-2$, it follows that $e'' \equiv 0$ and thus that e is a quadratic polynomial on $(0,1)$. On the other hand, we have seen that there exist at least $n+1$ points in $\Theta \cap (0,1)$, meaning that there are $n+1 > 1$ points in $(0,1)$ at which e' vanishes. We conclude that $e' \equiv 0$, i.e. e is a linear polynomial on $(0,1)$. It is easy to see that this is not optimal in terms of approximation.

Step 2. We claim that the $(n+2) \times (n+2)$-Vandermonde type matrix

$$V = \begin{pmatrix} s_0 & s_1 & \dots & s_{n+1} \\ 1 & 1 & \dots & 1 \\ s_0^2 & s_1^2 & \dots & s_{n+1}^2 \\ s_0^4 & s_1^4 & \dots & s_{n+1}^4 \\ \vdots & \vdots & \ddots & \vdots \\ s_0^{2n} & s_1^{2n} & \dots & s_{n+1}^{2n} \end{pmatrix}$$

is invertible for any distinct $n+2$ points $0 \leq s_0 < \dots < s_{n+1} \leq 1$. This is true by classical results of Lundengård (2017) for the $(n+1) \times (n+1)$ Vandermonde submatrix

$$V = \begin{pmatrix} 1 & \dots & 1 \\ s_0^2 & \dots & s_n^2 \\ \vdots & \ddots & \vdots \\ s_0^{2n} & \dots & s_n^{2n} \end{pmatrix}$$

since the points $s_0^2, \dots, s_n^2$ are distinct. It remains to show that the first row is linearly independent from the others, i.e. there exist no coefficients $a_0, \dots, a_k$ such that $s = \sum_{k=0}^n a_k s^{2k}$ at $(n+2)$ distinct points in $[0,1]$. Assume the contrary. Then there are $n+2$ distinct points $s_0 < \dots < s_{n+1} \in [0,1]$ such that

$$0 = s - \sum_{k=0}^n a_k s^{2k}, \qquad s \in \{s_0, \dots, s_{n+1}\}.$$

By Rolle's theorem, between two such points s_i, s_{i+1} there exists ξ_i such that

$$0 = \frac{d}{ds}\bigg|_{s=\xi_i} \left(s - \sum_{k=0}^{n} a_k s^{2k} \right).$$

The contradiction follows as in Step 1 of this proof.

Step 3. Combining the results of the second and third step of this proof, we can choose $\{s_0, \dots, s_{n+1}\} = \Theta$ and find a unique vector $\nu \in \mathbb{R}^{d+2}$ such that

$$V\nu = (1, 0, 0, \dots, 0)^T \tag{19}$$

The measure under consideration is now

$$\mu = \sum_{i=0}^{n+1} \mu_i \delta_{s_i} \quad \text{such that } \int_0^1 s^{2k} \,\mathrm{d}\mu_s = \begin{cases} (V\nu)_1 & k = 0 \\ (V\nu)_{k+2} & k \geq 1 \end{cases} = 0, \quad \int_0^1 s^{2k} \,\mathrm{d}\mu_s = (V\nu)_2 = 1$$

by construction. Thus the moment conditions are met. It remains to show that $\mu_i \cdot \big(s - \sum_{k=0}^{n} a_k s^{2k}\big)$ does not change sign in order to ensure that equality is attained in Hölder's inequality. Using Chebyshev's equi-oscillation theorem again, it suffices to show that μ_i and μ_{i+1} have opposite signs for all i.

For any $i \in \{0, \dots, n\}$, consider the unique even polynomial P of degree n such that $P(s_j) = 0$ for $0 \leq j \leq n+1$ except $j \in \{i, i+1\}$. Then, since P is an even polynomial of degree $\leq 2n$

$$0 = \int_0^1 P(s) \,\mathrm{d}\mu_s = \mu_i P(s_i) + \mu_{i+1} P(s_{i+1}), \tag{20}$$

but since P has $2n$ zeros at $\pm s_j$ for $i \notin \{i, i+1\}$, we find that $P(s) \neq 0$ for any $s \in [s_i, s_{i+1}]$. Thus $P(s_i)$ and $P(s_{i+1})$ have the same sign. In order to satisfy (20), we therefore find that μ_i and μ_{i+1} must have different signs. ∎

It remains to establish the claim in the proof of Theorem 8.

Proof To see this, we use the lower bound

$$\|\mu\| \geq \frac{1}{\mathrm{dist}_{L^\infty(\varepsilon,1)}\big(s,\, \mathrm{span}\{1, s^2, \dots, s^{2n}\}\big)} \tag{21}$$

from (17). By replacing the variable s by s^2, we find that

$$\mathrm{dist}_{L^\infty(\varepsilon,1)}\big(s,\, \mathrm{span}\{1, s^2, \dots, s^{2n}\}\big) = \mathrm{dist}_{L^\infty(\varepsilon^2,1)}\big(\sqrt{s},\, \mathrm{span}\{1, s, \dots, s^{n}\}\big).$$

Recall that the function $\sqrt{s}$ is an analytic function on the interval $[\varepsilon^2, 1]$ and

$$\sqrt{s} = 1 + \sum_{n=1}^{\infty} (-1)^{n+1} \frac{\prod_{k=1}^{n}(2k-1)}{2^n\, n!} (s-1)^n = 1 - \frac{1}{\sqrt{\pi}} \sum_{n=1}^{\infty} \frac{\Gamma(n+1/2)}{\Gamma(n+1)} (1-s)^n.$$

The coefficients decay asymptotically as $n^{-1/2}$ since

$$\lim_{n\to\infty} \left(\sqrt{n}\, \frac{\Gamma(n+1/2)}{\Gamma(n+1)} \right) = 1,$$

so for every $\delta > 0$, there exists $N \in \mathbb{N}$ which is independent of ε such that the L^∞-distance of the function $s \mapsto \sqrt{s}$ from the space $\mathcal{P}_n$ of polynomials of degree $\leq n$ is at most

$$\begin{aligned}\operatorname{dist}_{L^\infty(\varepsilon^2,1)}\big(\sqrt{s},\, \mathcal{P}_n\big) &\leq \max_{s\in[\varepsilon^2,1]} \left|\sqrt{s} - 1 - \frac{1}{\sqrt{\pi}} \sum_{k=1}^{n} \frac{\Gamma(k+1/2)}{\Gamma(k+1)} (1-s)^k\right| \\ &\leq \frac{1+\delta}{\sqrt{\pi n}} \sum_{k=n+1}^{\infty} (1-\varepsilon^2)^k \\ &= (1+\delta)\frac{(1-\varepsilon^2)^{n+1}}{\sqrt{\pi n}\,\varepsilon^2}.\end{aligned}$$

■

Appendix C. Brief Proofs of Known Results

In this appendix, we merely sketch the proofs of known results. For a more detailed introduction, we recommend e.g. (E and Wojtowytsch, 2020a). We begin by sketching a proof of Proposition 1, where we establish general properties of Barron functions.

Proof First claim. We note that, assuming existence of the integrals and for fixed $x \in \mathbb{R}^d$, we have

$$\begin{aligned} f_\pi(0) &= \int_{\mathbb{R}^{d+2}} a\,\sigma(b)\,\mathrm{d}\pi \\ |f_\pi(x) - f_\pi(0)| &= \left|\int_{\mathbb{R}^{d+2}} a\left\{\sigma(w^Tx+b) - \sigma(b)\right\}\mathrm{d}\pi\right| \leq \int_{\mathbb{R}^{d+2}} |a|\,|w^Tx|\,\mathrm{d}\pi \\ &\leq \frac{|x|}{2}\int_{\mathbb{R}^{d+2}} |a|^2 + |w|^2\,\mathrm{d}\pi. \end{aligned}$$

If the first integral exists, then also the integral defining $f_\pi(x)$ exists as the integrand is continuous and grows at most linearly. Then

$$f_\pi(x) = f_\pi(x) - f_\pi(0) + f_\pi(0) = \int_{\mathbb{R}^{d+2}} a\left\{\sigma(w^Tx+b) - \sigma(b)\right\}\mathrm{d}\pi + f_\pi(0).$$

Measurability is not an issue for fixed x due to the continuity of the integrand. For the sake of brevity, denote $h_{(a,w,b)}(x) = a\left\{\sigma(w^Tx+b) - \sigma(b)\right\}$. More generally, we note that the Bochner-integral

$$f = c + \int_{\mathbb{R}^{d+2}} \big(x \mapsto h_{(a,w,b)}(x)\big)\ \mathrm{d}\pi_{(a,w,b)}$$

converges in $C^0(K)$ for compact sets K and in $L^p(\mathbb{P})$ for $1 \leq p < \infty$ and probability distributions $\mathbb{P}$ with finite p-th moments in x, i.e. the function $(a,w,b) \mapsto h_{(a,w,b)}$ is Bochner integrable with respect to π when considered as a function with values in either $C^0(K)$ or $L^p(\mathbb{P})$. To see this, consider step functions

$$\tilde{h}_i = \sum_j 1_{Q_{ij}} h_{(a_{ij},w_{ij},b_{ij})}, \qquad \tilde{f}_i = \int_{\mathbb{R}^{d+2}} \tilde{h}_i\,\mathrm{d}\pi$$

where Q_{ij} are $(d+2)$-dimensional cubes of side length 2^{-i} whose union is $\bigcup_j Q_{ij} = [-2^i, 2^i]^{d+2}$ and $(a_{ij}, w_{ij}, b_{ij}) \in W_{ij}$. If $(a,w,b) \in Q_{ij}$, then

$$\begin{aligned}|h_{(a,w,b)}(x) - h_{(a_{ij},w_{ij},b_{ij})}(x)| &\leq |a - a_{ij}|\,|\sigma(w^T x + b)| + |a_{ij}|\,\big|\sigma(w^T x + b) - \sigma(w_{ij}^T x + b_{ij}\big| \\ &\leq |a - a_{ij}|\,|w^T x + b| + |a_{ij}|\big[|w - w_{ij}|\,|x| + |b - b_{ij}|\big] \\ &\leq C\left(\frac{|a-a_{ij}|^2 + |w-w_{ij}|^2 + |b-b_{ij}|^2}{\varepsilon} + \varepsilon\left\{|a|^2 + |w|^2|x|^2 + |b|^2\right\}\right)\end{aligned}$$

for any $\varepsilon > 0$. Fixing ε to be the square root of the side-length of Q_{ij}, we find that $\tilde{f}_i(x) \to f_\pi(x)$ pointwise for all x. Furthermore, $\tilde{f}_i$ is Lipschitz continuous in x uniformly in i, so $\tilde{f}_i$ converges to a limit in $C^0(K)$ by the compact embedding of Lipschitz functions in C^0, which coincides with the pointwise limit f_π. In other words, the Bochner integral exists in C^0. The argument follows in $L^p(\mathbb{P})$ by the dominated convergence theorem considering $|\tilde{f}_i|(x) \leq 2(1 + [f]_{\mathcal{B}}|x|)$ for all $x \in \mathbb{R}^d$.

Second claim. In this step, we show that V_0 is a Banach space and illustrate that $\mathcal{B}$ and $\mathcal{B}_0$ are different spaces. The fact that V_0 is a Banach space follows as (Siegel and Xu, 2021a, Lemma 1) from the previous claim, where we have shown the existence of $f \in \mathcal{B}_0$ as a Bochner integral in $L^2(\mathbb{P})$, i.e. as a continuous convex combination not only pointwise, but in a function space.

To see that $\mathcal{B} \neq \mathcal{B}_0$, observe that any $f \in \mathcal{B}$ can be decomposed into a positively one-homogeneous and a bounded part due to (E and Wojtowytsch, 2020a, Corollary 5.3). On the other hand, in one dimension, the function $f(x) = \log(1+x^2)$ satisfies $f(0) = f'(0) = 0$ and has an integrable second derivative $f''(x) = 2\frac{1-x^2}{(1+x^2)^2}$. By Proposition 4, we find that $f \in \mathcal{B}_0$. Since f is not bounded but grows sub-linearly, we conclude that $\mathcal{B} \subseteq \mathcal{B}_0 \not\subseteq \mathcal{B}$. The first inclusion follows from the fact that $[f]_{\mathcal{B}} \leq \|f\|_{\mathcal{B}}$ as shown next.

Third claim. The claim that $[f]_{\mathcal{B}} \leq \|f\|_{\mathcal{B}}$ is self-evident by definition, as the full Barron norm also limits the magnitude of the bias.

Fourth claim. Finally, we note that $f \in \mathcal{B}_0$ is Lipschitz-continuous, since

$$\begin{aligned}|f_\pi(x) - f_\pi(x')| &= \left|\int_{\mathbb{R}^{d+2}} a\left[\sigma(w^T x + b) - \sigma(w^T x' + b)\right] \mathrm{d}\pi_{(a,w,b)}\right| \\ &\leq \int_{\mathbb{R}^{d+2}} |a|\,|w^T(x - x')|\,\mathrm{d}\pi_{(a,w,b)} \leq |x - x'| \int_{\mathbb{R}^{d+2}} |a|\,|w|\,\mathrm{d}\pi_{(a,w,b)}.\end{aligned}$$

Taking the infimum over π (and optionally noting that $2|a|\,|w| \leq |a|^2 + |w|^2$), we find that $|f(x) - f(x')| \leq [f]_{\mathcal{B}}|x - x'|$. ∎

Proposition 2 is proved in (E and Wojtowytsch, 2020a, Theorem 5.18) and Corollary 3 follows from it directly. Let us sketch how the structure of one-dimensional Barron functions described in Proposition 4 can be understood.

Proof Upper bound. Let $a \in \mathbb{R}$ and $f \in C^2(\mathbb{R})$ be such that $f'' \in L^1(\mathbb{R})$. Then for $x > a$ we have

$$\begin{aligned}f(x) &= f(a) + \int_a^x f'(s) \cdot 1\,\mathrm{d}s = f(a) + f'(a)\,(x-a) - \int_a^x f''(s)\,(s-x)\,\mathrm{d}s \\ &= f(a) + f'(a)\,\sigma(x-a) + \int_a^\infty f''(s)\,\sigma(x-s)\,\mathrm{d}s\end{aligned}$$

and for $x < a$

$$f(x) = f(a) - \int_x^a f'(s) \cdot 1 \, \mathrm{d}s = f(a) - f'(a)\,(a-x) - \int_a^x f''(s)\,(s-x)\,\mathrm{d}s$$
$$= f(a) - f'(a)\,\sigma(a-x) + \int_a^\infty f''(s)\,\sigma(x-s)\,\mathrm{d}s.$$

Noting that the σ terms in the first expression vanish for when $x < a$ and vice versa, we find that

$$f(x) = f(a) + f'(a)\big[\sigma(x-a) - \sigma(a-x)\big] + \int_{\mathbb{R}} f''(s)\big[\sigma(x-s)\,1_{(a,\infty)}(s) + 1_{(-\infty,a)}(s)\,\sigma(s-x)\big]\,\mathrm{d}s.$$

Consequently, $f = f_\mu$ for a measure

$$\mu = f'(a)\,\big[\delta_{(1,a)} - \delta_{(-1,a)}\big] + f''(b)\,\cdot\,\big[\mathcal{H}^1|_{\{w=1,b>a\}} + \mathcal{H}^1|_{\{w=-1,b<a\}}\big]$$

where δ denotes the atomic point measure of mass one and $\mathcal{H}^1$ denotes the one-dimensional Hausdorff measure, restricted to half-lines $\{w = 1, b > a\}$ and $\{w = -1, b < a\}$. and hence

$$[f]_{\mathcal{B}} = \inf_{f=f_\mu} \|\mu\|_{TV} \leq 2 \inf_{a\in\mathbb{R}} |f'(a)| + \int_{\mathbb{R}} |f''(s)|\,\mathrm{d}s.$$

By approximation, the same is true if $f \notin C^2$ and f'' is merely a measure.

Lower bound direction. The bound

$$[f]_{\mathcal{B}} \leq [f]_{Lip} = \sup_{a\in\mathbb{R}} \max_{v\in\partial f(a)} |v| = \sup_{a\in\mathbb{R}} |f'(a)|$$

follows from Proposition 1 and the Rademacher Theorem on the differentiability of Lipschitz functions. For the second form of the lower bound, let $f \in \mathcal{B}_0$, i.e. there exists a measure μ on $\mathbb{R}^2$ such that

$$f(x) = \int_{\mathbb{R}^2} \sigma(w^T x + b)\mathrm{d}\mu_{(w,b)} = \int_{\{w=0\}} \sigma(b)\mathrm{d}\mu_{(w,b)} + \int_{\mathbb{R}^2} |w|\,\sigma\big(w/|w|\,x + b/|w|\big)\,\mathrm{d}\mu_{(w,b)}.$$

The second expression can be written as

$$f(x) = c + f^+(x) + f^-(x) = c + \int_{\mathbb{R}} \sigma(x+b)\,\mathrm{d}\mu_b^+ + \int_{\mathbb{R}} \sigma(-x+b)\,\mathrm{d}\mu_b^+$$

where

$$\mu_\pm = \psi_\sharp\big(|w| \cdot 1_{\{\pm w>0\}} \cdot \mu\big), \qquad \psi(w,b) = b/|w|,$$

i.e. $\mu_\pm$ is the push-forward of the measure which has density $|w|$ with respect to μ onto the real line. If $\phi \in C_c^\infty(\mathbb{R})$ is any function, then by exchanging the order of integration and integrating by parts, we find that

$$\begin{aligned}\int_{-\infty}^\infty f^+(x)\,\phi''(x)\,\mathrm{d}x &= \int_{-\infty}^\infty \phi''(x) \int_{\mathbb{R}} \sigma(x+b)\,\mathrm{d}\mu_b^+\,\mathrm{d}x \\ &= \int_{\mathbb{R}}\int_{-b}^\infty \phi''(x)\,(x+b)\,\mathrm{d}x\,\mathrm{d}\mu_b^+ \\ &= -\int_{\mathbb{R}}\int_{-b}^\infty \phi'(x)\,\mathrm{d}x\,\mathrm{d}\mu_b^+ \\ &= \int_{\mathbb{R}} \phi(b)\,\mathrm{d}\mu_b^+\end{aligned}$$

we find that $(f^+)'' = \mu^+$ in the distributional sense, and thus $f'' = \mu^+ + \mu^-$. In particular,

$$\|f''\|_{TV} = \|\mu^+ + \mu^-\|_{TV} \leq \inf_{\mu} \|\mu^+\|_{TV} + \|\mu^-\|_{TV} \leq \inf_{\mu} \int_{\mathbb{R}^2} |w| \,\mathrm{d}|\mu|_{(w,b)} = [f]_{\mathcal{B}}.$$

■

We sketch a proof of the direct approximation theorem for Barron spaces (Proposition 5).

Proof Step 1. Consider the Hilbert space $L^2(\mathbb{P})$ and observe that $h_{(a,w,b)} \in H$ defined by $h_{(a,w,b)}(x) = a\left\{\sigma(w^T x + b) - \sigma(b)\right\}$ has norm at most

$$\|h_{(a,w,b)}\|_H^2 = \int_{\mathbb{R}^d} a^2 \left[\sigma(w^T x + b) - \sigma(b)\right]^2 \mathrm{d}\mathbb{P} \leq a^2 \int_{\mathbb{R}^d} |w^T x|^2 \,\mathrm{d}\mathbb{P}.$$

We use Proposition 1 to write $f \in \mathcal{B}_0$ as

$$f(x) = f(0) + \int_{\mathbb{R}^{d+2}} h_{(a,w,b)}(x) \,\mathrm{d}\pi_{(a,w,b)}.$$

Step 2. Using the homogeneity relation $\sigma(z) = \lambda^{-1}\sigma(\lambda z)$, the distribution π can be normalized such that

$$|a|^2 = |w|^2 = \frac{1}{2} \int_{\mathbb{R}^{d+2}} |a'|^2 + |w'|^2 \,\mathrm{d}\pi_{(a',w',b')}$$

almost surely by considering the push-forward of π along the map

$$T : \mathbb{R}^{d+2} \to \mathbb{R}^{d+2}, \qquad T(a,w,b) = \left(a\sqrt{\frac{|w|}{|a|}},\, w\sqrt{\frac{|a|}{|w|}}, w\sqrt{\frac{|a|}{|w|}}\right)$$

if $a, w \neq 0$ and $T(a,w,b) = 0$ otherwise, which satisfies $f_{T_\sharp \pi} \equiv f_\pi$. Thus for any $\varepsilon > 0$, $f - f(0)$ is in the H-closed convex hull of the family

$$\mathcal{G}_{\|f\|_{\mathcal{B}}+\varepsilon} = \{h_{(a,w,b)} : |a| = |w| \leq \|f\|_{\mathcal{B}} + \varepsilon\}.$$

Step 3. By the Maurey-Barron-Jones Lemma (Barron, 1993, Lemma 1), for every $m \in \mathbb{N}$ and every $\varepsilon' > 0$, there exist $h_{(a_i,w_i,b_i)} \in \mathcal{G}_{\|f\|_{\mathcal{B}}+\varepsilon}$ such that

$$\left\| f - f(0) - \frac{1}{m} \sum_{i=1}^m h_{(a_i,w_i,b_i)} \right\|_H \leq \frac{\|f\|_{\mathcal{B}} + \varepsilon}{\sqrt{m}} + \varepsilon'.$$

As the vectors (a_i, w_i, b_i) are constrained to a compact domain of $\mathbb{R}^{d+2}$ and the map $\mathbb{R}^{d+2} \to H$, $(a,w,b) \mapsto h_{(a,w,b)}$ is continuous, we can set $\varepsilon, \varepsilon' \to 0$ and obtain the result without constant by an appropriate subsequence.

Finally, we write $c = f(0) + \frac{1}{m}\sum_{i=1}^m a_i \sigma(b_i)$ for compatibility with the original notation.

■

Appendix D. Further Results

D.1 On the Decay of $f_d^*(x)$ for $x \neq 0$

Numerical experiments in Appendix A suggest that $f_d^*(x)$ decays to zero exponentially fast for $x \neq 0$. While we cannot prove this in full generality, we show that

$$0 \leq f_d^*(x) \leq C\, d^{3/2} \left(\frac{1-|x|^2}{|x|}\right)^{\frac{d-3}{2}}$$

for a constant $C > 0$ which is independent of d. In particular, $f_d^*(x) \to 0$ exponentially fast in d if $|x| > 0.62$. To see this, observe that

$$\begin{aligned} f_d^*(r) &= c_d \int_{-1}^{1} g(rs)\,(1-s^2)^{\frac{d-3}{2}}\,\mathrm{d}s = 2c_d\, r^{\frac{1-d}{2}} \int_0^r g(z)(r^2-z^2)^{\frac{d-3}{2}}\,\mathrm{d}z \\ &= -2c_d\, r^{\frac{1-d}{2}} \int_r^1 g(z)(r^2-z^2)^{\frac{d-3}{2}}\,\mathrm{d}z \end{aligned}$$

for $r < 1$ since g is $L^2(0,1)$-orthogonal to the polynomial $(r^2-z^2)^{\frac{d-3}{2}}$. Since $\|g\|_{L^\infty(0,1)} \leq \gamma_{\frac{d-1}{2}}$, we may estimate

$$|f_d^*(r)| \leq 2c_d\, r^{\frac{1-d}{2}}\, \gamma_d\, (1-r^2)^{\frac{d-3}{2}} = \frac{2c_d \gamma_{\frac{d-1}{2}}}{r} \left(\frac{1-r^2}{r}\right)^{\frac{d-3}{2}}.$$

The pre-factor grows as $d^{3/2}$ since $\gamma_d \sim d$ and

$$c_d = \frac{1}{\int_{-1}^1 (1-s^2)^{\frac{d-3}{2}}\,\mathrm{d}s} = \frac{\Gamma\left(d/2\right)}{\sqrt{\pi}\,\Gamma\left(\frac{d-1}{2}\right)} \sim \sqrt{\frac{d}{2\pi}}.$$

Finally, we note that $(1-r^2)/r < 1$ holds for positive r if and only if $r > \frac{\sqrt{5}-1}{2} \approx 0.618$.

D.2 Non-radial Minimum Norm Interpolants

In this note, we constructed

$$f_d^* \in \operatorname*{argmin}_{f \in \mathcal{F}} [f]_{\mathcal{B}}, \qquad \mathcal{F} = \left\{ f \in \mathcal{B}_0(\mathbb{R}^d) : f(0) = 1 \text{ and } f \equiv 0 \text{ on } \mathbb{R}^d \setminus B_1(0) \right\}. \tag{22}$$

Since both the Barron semi-norm and the class $\mathcal{F}$ are convex and invariant under rotations of the data domain, we find that there exists at least one minimizer which is radially symmetric. By direct construction, we saw that this minimizer

$$f_d^*(x) = 1 + \sum_{i=0}^{\frac{d+1}{2}} \mu_i \fint_{S^{d-1}} \sigma(\nu^T x - b_i)\,\mathrm{d}\mathcal{H}_\nu^{d-1}$$

is unique, at least if d is odd. The biases $0 = b_0 < \cdots < b_{(d+1)/2} = 1$ and weights $\mu_i \neq 0$ are given by the optimization process. Our proof does not exclude the existence of other

minimizers, which are not radially symmetric. In fact, assume that $\phi_i \in L^\infty(S^{d-1})$ for $i = 0, \dots, \frac{d+1}{2}$ are functions such that

$$f_\phi(x) = \sum_{i=0}^{\frac{d+1}{2}} \fint_{S^{d-1}} \sigma(\nu^T x - b_i)\, \phi_i(\nu)\, \mathrm{d}\mathcal{H}^{d-1}_\nu = 0 \qquad \forall\ |x| \geq 1. \tag{23}$$

Then trivially also $f_\phi(0) = 0$ since $b_i \geq 0$, and thus

$$(f_d^* + \varepsilon f_\phi)(x) = 1 + \sum_{i=0}^{\frac{d+1}{2}} \fint_{S^{d-1}} \big(\mu_i + \varepsilon\phi_i(\nu)\big)\, \sigma(\nu^T x - b_i)\, \mathrm{d}\mathcal{H}^{d-1}_\nu = \begin{cases} 1 & x = 0 \\ 0 & |x| \geq 1 \end{cases}.$$

Since f_d^* is the *unique* radial solution, we can average in the radial direction and observe that $\int_{S^{d-1}} \phi_i(\nu) = 0$ for all $i = 0, \dots, \frac{d+1}{2}$. The Barron norm of the combined solution is

$$\sum_{i=0}^{(d+1)/2} \frac{\|\mu_i + \varepsilon\phi_i\|_{L^1(S^{d-1})}}{\mathcal{H}^{d-1}(S^{d-1})} = \sum_{i=0}^{(d+1)/2} |\mu_i|$$

if ε is so small that $\varepsilon\|\phi_i\|_{L^\infty} \leq |\mu_i|$ for all i, since the function $\mu_i + \varepsilon\phi_i$ does not change signs in this case, and the integral of ϕ_i averages to zero. In particular, if $(\phi_0, \dots, \phi_{(d+1)/2})$ exist such that f_ϕ is supported in $\overline{B_1(0)}$ and fails to be radial, then a non-radial minimizer exists.

By considering the behavior of f_ϕ at infinity, we establish two conditions: $\sum_{i=0}^{(d+1)/2} \phi_i \equiv 0$ in order to have f_ϕ bounded, and $\sum_{i=0}^{(d+1)/2} b_i\phi_i \equiv 0$ in order to have $\lim_{x\to\infty} f_\phi(x) = 0$.

Lemma 13. *Assume there exist $\frac{d+3}{2}$ measures $\bar\mu_i$ on S^{d-1} such that*

$$f_{\bar\mu}(x) := \sum_{i=0}^{\frac{d+1}{2}} \fint_{S^{d-1}} \sigma(\nu^T x - b_i)\, \mathrm{d}\bar\mu_i = 0$$

for all $|x| \geq 1$ and $f_{\bar\mu}(x) \not\equiv 0$. Then there exists a minimizer $\hat f_d \in \mathcal{F}$ of the Barron seminorm which is not radially symmetric. Without loss of generality, we may assume that $\hat f_d$ is radially symmetric with respect to $(x_2, \dots, x_d)$.

Proof Step 1. Assume for now that $f_{\bar\mu}$ is identically zero. Let ψ_δ be a C^∞-probability density on the group of rotations $SO(d)$ which is supported in an δ-neighbourhood of the unit matrix, and let H be the Haar measure on $SO(d)$. Define the radial mollification

$$\begin{aligned} f_{\bar\mu,\delta}(x) &= \int_{SO(d)} \psi_\delta(O)\, f_{\bar\mu}(O^T x)\, \mathrm{d}H_O \\ &= \sum_{i=0}^{\frac{d+1}{2}} \fint_{S^{d-1}} \left(\int_{SO(d)} \psi_\delta(O)\, \sigma((O\nu)^T x - b_i)\, \mathrm{d}H_O \right) \mathrm{d}\bar\mu_{i,\nu} \\ &= \sum_{i=0}^{\frac{d+1}{2}} \int_{S^{d-1}} \sigma(\nu^T x - b_i)\, \mathrm{d}\tilde\mu_i \end{aligned}$$

where

$$\tilde{\mu}_{i,\delta}(B) = \int_{SO(d)} \psi_\delta(O)\, \bar{\mu}_i(O \cdot B)\, \mathrm{d}H_O.$$

We make three observations.

1. $f_{\bar{\mu},\delta}(x) = 0$ if $x = 0$ or $|x| \geq 1$.

2. $f_{\bar{\mu},\delta} \to f_{\bar{\mu}}$ as $\delta \to 0$ (pointwise and locally uniformly), so $f_{\bar{\mu},\delta}$ cannot be identically zero for sufficiently small $\delta > 0$.

3. $\tilde{\mu}_i$ is absolutely continuous with respect to the uniform distribution on the sphere since

$$|\tilde{\mu}_i|(B) \leq \|\psi_\delta\|_{L^\infty} \|\bar{\mu}_i\|_{TV}.$$

   Due to the uniform estimate, the Radon-Nikodym derivative $\phi_{i,\delta} := \frac{\mathrm{d}\tilde{\mu}_{i,\delta}}{\mathrm{d}\mathcal{H}^{d-1}}$ is an $L^\infty(S^{d-1})$-function.

We now fix ε, δ small enough, write $\phi_i = \phi_{i,\delta}$ and note that $f_d^* + \varepsilon f_\phi$ is also a solution to (22). In particular, f_ϕ cannot be radially symmetric since f_d^* is the unique radially symmetric minimizer.

Step 2. Take f_ϕ to be non-trivial as implied by step 1. Then there exists at least one direction $\bar{\nu}$ such that $f_\phi(t\bar{\nu}) \not\equiv 0$. Without loss of generality, we may take $\bar{\nu} = e_1$. We can now average over all rotations which leave e_1 fixed. The resulting function $\hat{f}_\phi$ is radially symmetric in all components orthogonal to e_1, i.e. in $(x_2, \ldots, x_d)$. Since we only average over rotations which leave the e_1-direction fixed, we have $\hat{f}_\phi(te_1) = f_\phi(te_1) \not\equiv 0$. In particular, we may assume that f_ϕ has the desired symmetry. ∎

The question whether there exists $\bar{\mu} = (\bar{\mu}_0, \ldots, \bar{\mu}_{(d+1)/2})$ such that $f_{\bar{\mu}} \equiv 0$ on $\mathbb{R}^d \setminus B_1(0)$ but $f_{\bar{\mu}} \not\equiv 0$ on $\mathbb{R}^d$ can be rephrased in terms of functional analysis. Namely, if we understand $\bar{\mu}$ as an element of the dual space Z^* of $Z := C^0(S^{d-1}; \mathbb{R}^{(d+3)/2})$ and we associate to $x \in \mathbb{R}^d$ the function $h_x \in Z$ given by $\nu \mapsto \big(\sigma(\nu^T x - b_0), \ldots, \sigma(\nu^T x - b_{(d+1)/2})\big)$, then we can write $f_{\bar{\mu}}(x) = \langle \bar{\mu}, h_x \rangle_{Z^*,Z}$ as a duality product.

In particular, we consider two subspaces $V_1, V_2 \subseteq Z$:

$$V_1 = \mathrm{span}\{h_x : x \in \mathbb{R}^d\}, \qquad V_2 = \mathrm{span}\{h_x : |x| \geq 1\}. \tag{24}$$

Obviously $V_2 \subseteq V_1$. We note the following: If $\overline{V_2} \neq \overline{V_1}$, then by the Hahn-Banach theorem there exists $\mu \in Z^*$ such that $\langle \mu, v \rangle = 0$ for all $v \in \overline{V_2}$ but not all $v \in \overline{V_1}$. It is easy to see by contradiction that there exists in particular h_x with $|x| < 1$ such that $f_\mu(x) = \langle \mu, h_x \rangle_{Z^*,Z} \neq 0$. Note that $x \neq 0$ since $f_\mu(0) = 0$ for any μ by design.

We have thus proved the following.

Corollary 14. *Denote $Z := C^0(S^{d-1}; \mathbb{R}^{(d+3)/2})$ and $h_x \in Z$ defined by $h_x(\nu) = \big(\sigma(\nu^T x - b_0), \ldots, \sigma(\nu^T x - b_{(d+1)/2})\big)$. Consider the subspaces V_1, V_2 of Z as in (24). There exists a non-radial solution f of the minimization problem (22) if and only if $\overline{V_1} \neq \overline{V_2}$.*

References

Navid Ardeshir, Daniel J Hsu, and Clayton H Sanford. Intrinsic dimensionality and generalization properties of the r-norm inductive bias. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3264–3303. PMLR, 2023.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48), 2020.

Serge Bernstein. *Sur l'ordre de la meilleure approximation des fonctions continues par des polynômes de degré donné*, volume 4. Hayez, imprimeur des académies royales, 1912.

Etienne Boursier and Nicolas Flammarion. Penalising the biases in norm regularisation enforces sparsity. *arXiv preprint arXiv:2303.01353*, 2023.

Sébastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. *Advances in Neural Information Processing Systems*, 34:28811–28822, 2021.

Yurii D Burago and Viktor A Zalgaller. *Geometric inequalities*, volume 285. Springer Science & Business Media, 2013.

Andrei Caragea, Philipp Petersen, and Felix Voigtlaender. Neural network approximation and estimation of classifiers with classification boundary in a Barron class. *arXiv:2011.09363 [math.FA]*, 2020.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. *arxiv:2002.04486 [math.OC]*, 2020.

Manfred Dobrowolski. *Angewandte Funktionalanalysis: Funktionalanalysis, Sobolev-Räume und elliptische Differentialgleichungen*. Springer-Verlag, 2010.

Weinan E and Stephan Wojtowytsch. Representation formulas and pointwise properties for Barron functions. *Calc. Var. Partial Differential Equations*, 61(46), 2020a.

Weinan E and Stephan Wojtowytsch. On the Banach spaces associated with multi-layer ReLU networks of infinite width. *CSIAM Trans. Appl. Math.*, 1(3):387–440, 2020b.

Weinan E and Stephan Wojtowytsch. Kolmogorov width decay and poor approximators in machine learning: Shallow neural networks, random feature models and neural tangent kernels. *Res Math Sci*, 8(5), 2021.

Weinan E and Stephan Wojtowytsch. On the emergence of simplex symmetry in the final and penultimate layers of neural network classifiers. In *Mathematical and Scientific Machine Learning*, pages 270–290. PMLR, 2022.

Weinan E, Chao Ma, and Qingcan Wang. A priori estimates of the population risk for residual networks. *arXiv:1903.02154 [cs.LG]*, 2019a.

Weinan E, Chao Ma, and Lei Wu. A priori estimates of the population risk for two-layer neural networks. *Comm. Math. Sci.*, 17(5):1407 – 1425, 2019b.

Weinan E, Chao Ma, and Lei Wu. Machine learning from a continuous viewpoint. *arxiv:1912.12777 [math.NA]*, 2019c.

Weinan E, Chao Ma, and Lei Wu. The Barron space and the flow-induced function spaces for neural network models. *arXiv:1906.08039 [cs.LG]*, 2019d.

Weinan E, Chao Ma, Stephan Wojtowytsch, and Lei Wu. Towards a mathematical understanding of neural network-based machine learning: what we know and what we don't. *CSIAM Trans. Appl. Math.*, 1(4):561–615, 2020.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on learning theory*, pages 907–940, 2016.

Rémi Gribonval, Gitta Kutyniok, Morten Nielsen, and Felix Voigtlaender. Approximation spaces of deep neural networks. *Constructive approximation*, 55(1):259–367, 2022.

Boris Hanin. Ridgeless interpolation with shallow relu networks in $1d$ is nearest neighbor curvature extrapolation and provably generalizes on lipschitz functions. *arXiv preprint arXiv:2109.12960*, 2021.

Juncai He, Lin Li, Jinchao Xu, and Chunyue Zheng. Relu deep neural networks and linear finite elements. *arXiv preprint arXiv:1807.03973*, 2018.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

David Kincaid, David Ronald Kincaid, and Elliott Ward Cheney. *Numerical analysis: mathematics of scientific computing*, volume 2. American Mathematical Soc., 2009.

Jason M Klusowski and Andrew R Barron. Approximation by combinations of relu and squared relu ridge functions with ℓ^1 and ℓ^0 controls. *IEEE Transactions on Information Theory*, 64(12):7649–7656, 2018.

Guy Kornowski, Gilad Yehudai, and Ohad Shamir. From tempered to benign overfitting in relu neural networks. *arXiv preprint arXiv:2305.15141*, 2023.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Zhong Li, Chao Ma, and Lei Wu. Complexity measures for neural networks with general activation functions using path-based norms. *arXiv preprint arXiv:2009.06132*, 2020.

B Llanas and FJ Sainz. Constructive approximate interpolation by neural networks. *Journal of Computational and Applied Mathematics*, 188(2):283–308, 2006.

Christos Louizos, Max Welling, and Diederik P Kingma. Learning sparse neural networks through l_0 regularization. *arXiv preprint arXiv:1712.01312*, 2017.

Zhou Lu. A note on the representation power of GHHs. *arXiv preprint arXiv:2101.11286*, 2021.

Karl Lundengård. *Generalized Vandermonde matrices and determinants in electromagnetic compatibility*. PhD thesis, Mälardalen University, 2017.

Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width ReLU nets: The multivariate case. *arXiv preprint arXiv:1910.01635*, 2019.

Rahul Parhi and Robert D Nowak. Banach space representer theorems for neural networks and ridge splines. *J. Mach. Learn. Res.*, 22:43–1, 2021.

Rahul Parhi and Robert D Nowak. What kinds of functions do deep neural networks learn? insights from variational spline theory. *SIAM Journal on Mathematics of Data Science*, 4(2):464–489, 2022.

Jiyoung Park, Ian Pelakh, and Stephan Wojtowytsch. Minimum norm interpolation by perceptra: Explicit regularization and implicit bias. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Itay Safran and Jason Lee. Optimization-based separations for neural networks. In *Conference on Learning Theory*, pages 3–64. PMLR, 2022.

Itay Safran and Ohad Shamir. Depth-width tradeoffs in approximating natural functions with neural networks. In *International conference on machine learning*, pages 2979–2987. PMLR, 2017.

Itay Safran, Ronen Eldan, and Ohad Shamir. Depth separations in neural networks: what is actually being separated? In *Conference on Learning Theory*, pages 2664–2666. PMLR, 2019.

Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In *Conference on Learning Theory*, pages 2667–2690. PMLR, 2019.

Jonathan W Siegel and Jinchao Xu. On the approximation properties of neural networks. *arXiv preprint arXiv:1904.02311*, 2019.

Jonathan W Siegel and Jinchao Xu. Approximation rates for neural networks with general activation functions. *Neural Networks*, 128:313–321, 2020.

Jonathan W Siegel and Jinchao Xu. Characterization of the variation spaces corresponding to shallow neural networks. *arXiv preprint arXiv:2106.15002*, 2021a.

Jonathan W Siegel and Jinchao Xu. Optimal approximation rates and metric entropy of relu^k and cosine networks. *arXiv preprint arXiv:2101.12365*, 2021b.

Jonathan W Siegel and Jinchao Xu. Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks. *arXiv preprint arXiv:2101.12365*, 2021c.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Suraj Srinivas, Akshayvarun Subramanya, and R Venkatesh Babu. Training sparse neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 138–145, 2017.

Matus Telgarsky. Benefits of depth in neural networks. In *Conference on learning theory*, pages 1517–1539. PMLR, 2016.

Lloyd N Trefethen. *Approximation Theory and Approximation Practice, Extended Edition*. SIAM, 2019.

Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, et al. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873): 590–596, 2021.

Richard S Varga and Amos J Carpenter. On the Bernstein conjecture in approximation theory. *Constructive Approximation*, 1(1):333–348, 1985.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Luca Venturi, Samy Jelassi, Tristan Ozuch, and Joan Bruna. Depth separation beyond radial functions. *The Journal of Machine Learning Research*, 23(1):5309–5364, 2022.

Yongji Wang, Ching-Yao Lai, Javier Gómez-Serrano, and Tristan Buckmaster. Self-similar blow-up profile for the Boussinesq equations via a physics-informed neural network. *arXiv preprint arXiv:2201.06780*, 2022.

Stephan Wojtowytsch. https://colab.research.google.com/drive/1ofrzlafdq73ev1-mgmmnuua0lft0f5gg?usp=sharing, 2022.