Proceedings of Machine Learning Research vol 65:1–26, 2017

On the Ability of Neural Nets to Express Distributions

Holden Lee HOLDENL@PRINCETON.EDU
Princeton University, Mathematics Department

Rong Ge RONGGE@CS.DUKE.EDU
Duke University, Computer Science Department

Tengyu Ma TENGYU@CS.PRINCETON.EDU
Andrej Risteski RISTESKI@PRINCETON.EDU
Sanjeev Arora ARORA@CS.PRINCETON.EDU
Princeton University, Computer Science Department

Abstract

Deep neural nets have caused a revolution in many classification tasks. A related ongoing revolution—also theoretically not understood—concerns their ability to serve as generative models for complicated types of data such as images and texts. These models are trained using ideas like variational autoencoders and Generative Adversarial Networks.

We take a first cut at explaining the expressivity of multilayer nets by giving a sufficient criterion for a function to be approximable by a neural network with n hidden layers. A key ingredient is Barron's Theorem (Barron, 1993), which gives a Fourier criterion for approximability of a function by a neural network with 1 hidden layer. We show that a composition of n functions which satisfy certain Fourier conditions ("Barron functions") can be approximated by a $n+1$-layer neural network.

For probability distributions, this translates into a criterion for a probability distribution to be approximable in Wasserstein distance—a natural metric on probability distributions—by a neural network applied to a fixed base distribution (e.g., multivariate gaussian).

Building up recent lower bound work, we also give an example function that shows that composition of Barron functions is more expressive than Barron functions alone.

Keywords: neural network, generative model, function approximation, Fourier transform

1. Introduction

Deep neural networks have led to state-of-the-art performance on classification tasks in many domains such as computer vision, speech recognition, and reinforcement learning (Bengio et al., 2013; Schmidhuber, 2015). One can view a neural network as a way to learn a function mapping inputs x to outputs y. For image classification, the input is a vector representing an image and the output can be probabilities of being in various classes.

But another recent (and less understood) use of neural networks is as generative models for complicated probability distributions, such as distributions over images on ImageNet, handwritten characters from various alphabets, or speech. Here the network may map a stochastic input—such as a uniform normal gaussian—to a realistic image. Such networks are trained using various methods such as variational autoencoders (Kingma and Welling (2013), Rezende et al. (2014)) or generative adversarial networks (GANs) (Goodfellow et al. (2014)). A GAN consists of a repeated zero-sum game between two networks: the *generator* attempts to imitate a given probability distribution; it

© 2017 H. Lee, R. Ge, T. Ma, A. Risteski & S. Arora.

obtains its samples by passing a base distribution (e.g. a gaussian) through its neural network. The *discriminator* attempts to distinguish between samples from the generator and the true distribution, and thus forces the generator to improve over many repetitions.

The current paper is concerned with the following natural question that appears not to have been studied before: Why are deep neural networks so well-suited to efficiently generate many distributions that occur in nature?

1.1. Our work

We give a sufficient criterion for a function to be approximable by a neural network with n hidden layers (Theorem 3.1). This criterion holds with respect to any distribution of inputs supported on a compact set. As a consequence of our main result, we obtain a criterion for a distribution to be approximately generated by a neural network with n hidden layers in the Wasserstein metric W_2, a natural metric on the space of distributions (Corollary 3.3).

Our criterion relies on Fourier properties of the function. We build on Barron's Theorem Barron (1993), which says that if a certain quantity involving the Fourier transform is small, then the function can be approximated by a neural network with one hidden layer and a small number of nodes. Calling such a function a Barron function, our criterion roughly says that if a distribution is generated by a composition of n Barron functions, then the distribution can be approximately generated by a neural network with n hidden layers.

Many nice functions, such as polynomials and ridge functions, are Barron; this property is also preserved under natural operations such as linear combinations. Thus, our result says that if nature creates a distribution by starting from a base distribution (such as a gaussian) and applying a sequence of functions in this class, then we can also generate that distribution with a neural network.

This "correspondence" between compositions of Barron functions and multi-layer neural networks raises questions analogous to those raised about neural nets: for example, are compositions of k Barron functions more expressive than Barron functions? Using a technique to lower-bound the Barron constant (Theorem 4.2), we show a separation theorem between Barron functions and composition of Barron functions (Theorem 4.1). This parallels —and is inspired by—the separation between 2-layer and 3-layer neural networks in Eldan and Shamir (2015).

1.2. Related work

Despite the practical success of neural networks, we lack a good theoretical understanding of their effectiveness. An initial attempt to understand the effectiveness of neural networks was by their function approximation properties. A series of works showed that any continuous function in a bounded domain can be approximated by a sufficiently large 2-layer neural network (Cybenko (1989), Funahashi (1989), Hornik et al. (1989)). However, the network size can be exponential in the dimension. Barron (Barron (1993)) gave a upper bound for the size of the network required in terms of a Fourier criterion. He showed that a function f can be approximated in L^2 up to error ε by a 2-layer neural network with $O\left(\frac{C_f^2}{\varepsilon}\right)$ units, where C_f depends on Fourier properties of f. One remarkable consequence is that representationally speaking, neural nets can evade the curse of dimensionality: the number of parameters required to obtain a fixed error increases linearly, rather than superlinearly, in the number of dimensions. (Fixing the number of nodes in the hidden layer, the number of parameters scales linearly in the number of dimensions.)

However, such approximability results only explain a small part of the success of neural networks. Firstly, they only deal with 2-layer neural networks. Empirically speaking, deep neural networks—networks with many layers—appear to be much more effective than shallow neural networks. There have been several attempts to explain the effectiveness of deep neural networks. Following the paradigm in circuit complexity, one produces a function f that can be computed by a deep neural network but requires exponentially many nodes to be computed by a shallow neural network. Eldan and Shamir (Eldan and Shamir (2015)) show a certain radial function can be approximated by a 3-layer neural net but not by a 2-layer neural net with a subexponential number of nodes. Daniely (2017) shows such a separation but with respect to the uniform distribution on the sphere. Telgarsky (Telgarsky (2016)) shows such a separation between k^2-layer and k-layer neural networks. Cohen, Sharir, and Shashua (Cohen et al. (2015)) show a separation for a different model, a certain type of convolutional neural net architecture. Kane and Williams (Kane and Williams (2016)) show super-linear gate and super-quadratic wire lower bounds for depth-two and depth-three threshold circuits, which can be thought of as a boolean analogue to neural networks.

Secondly, these works—as well as our paper—do not address how to learn neural networks, or why the established method, gradient descent, has been so successful. Barron (1993) and Barron (1994) address the generalization theory, and show that the nodes can be chosen "greedily"; however the optimization problem is nonconvex. Under the assumption that certain properties of the input distribution (related to the score function) are known and that the function is exactly representable by a 2-layer neural network, Janzamin, Sedghi, and Anandkumar (Janzamin et al. (2015)) give an algorithm inspired by Barron's Fourier criterion and utilizing tensor decomposition, to learn 2-layer neural networks.

Finally, we note that the learnability for distributions has been studied for discrete distributions (Kearns et al., 1994).

Organization of the paper We explain Barron's original theorem in Section 2, our criterion for representation by multi-layer neural networks in Section 3, and give our separation result in Section 4. Most proofs and background on Fourier analysis are left in Appendix.

1.3. Notation and Definitions

First, we formally define the model of a feedforward neural network that we will use.

Definition 1.1 *A* ***neural network with*** n ***hidden layers*** *(also referred to as a* $n+1$*-layer neural network) has an associated input space* $\mathbb{R}^{m_0}$*, output space* $\mathbb{R}^{m_{n+1}}$*, and* n *hidden layers of sizes* $m_1, \ldots, m_n \in \mathbb{N}$*.It has parameters* $A^{(l)} \in \mathbb{R}^{m_{l-1}\times m_l}$ *and* $b^{(l)} \in \mathbb{R}^{m_l}$ *for* $1 \le l \le n+1$*. The neural network has a fixed activation function* $\sigma : \mathbb{R} \to \mathbb{R}$*, and if* x *is a vector then* $\sigma(x)$ *denotes componentwise application. On input* $x \in \mathbb{R}^{m_0}$*, the network computes*

$$x^{(0)} := x \tag{1}$$

$$x^{(l)} := \sigma(A^{(l-1)}x^{(l-1)} + b^{(l)}) \qquad 1 \le l \le n \tag{2}$$

$$x^{(n+1)} := A^{(n+1)}x^{(n)} + b^{(n+1)}. \tag{3}$$

and outputs $x^{(n+1)}$*. This can also be written out in terms of the components:*

$$x_j^{(l)} := \sigma\left(\sum_{k=1}^{m_l} A_{jk}^{(l-1)} x_k^{(l-1)} + b_k^{(l-1)}\right).$$

Common choices of activation functions σ include the logistic function $\frac{1}{1+e^{-x}}$, $\tanh(x)$, and the ReLU function $\max\{0, x\}$.

Definition 1.2 *For a function $f\colon \mathbb{R}^m \to \mathbb{R}^n$, define $\mathrm{Lip}(f) = \mathrm{Lip}_2(f)$, the Lipschitz constant of f with respect to the L^2 norm, by*

$$\inf \left\{C : \forall x, y, \|f(x) - f(y)\|_2 \le C \|x - y\|_2 \right\}.$$

Let B_n be the unit ball in n dimensions$\{x \in \mathbb{R}^n : \|x\| \le 1\}$. For sets A, B and a scalar r, let

$$A + B := \{x + y : x \in A, y \in B\}, \quad rA := \{rx : x \in A\}. \tag{4}$$

For example, rB_n denotes the ball of radius r in n dimensions, and $A + rB_n$ is the neighborhood of radius r around A.

Let $\|\cdot\| = \|\cdot\|_2$ denote the usual Euclidean norm on vectors in $\mathbb{R}^n$. For a function f, let $f^\vee(x) := f(-x)$. (This notation is often used in Fourier analysis.) Let $f^{(n)}(x) = \frac{d^n}{dx^n} f(x)$ denote the nth derivative, and $\Delta f = \sum_{i=1}^n \frac{\partial^2}{\partial {x_i}^2} f$ denote the Laplacian.

2. Barron's Theorem

For $f \in L^1(\mathbb{R})$ we define the Fourier transform of $f\colon \mathbb{R}^n \to \mathbb{R}$ with the following normalization.

$$\widehat{f}(\omega) := \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} f(x) e^{-i\langle \omega, x\rangle} \, dx. \tag{5}$$

For vector-valued functions $f\colon \mathbb{R}^n \to \mathbb{R}^m$, define the Fourier transform componentwise.

The inverse Fourier transform is

$$(\mathcal{F}^{-1} g)(x) := \int_{\mathbb{R}^n} g(\omega) e^{i\langle \omega, x\rangle} \, dx = (2\pi)^n \widehat{g}^\vee$$

The Fourier inversion formula, which holds for all sufficiently "nice" functions, is

$$f(x) = \int_{\mathbb{R}^n} \widehat{f}(x) e^{i\langle \omega, x\rangle} \, dx. = (2\pi)^n \widehat{\widehat{f}}^\vee$$

For background on Fourier analysis with rigorous statements, see Appendix A.

Barron (1993) defines a norm on functions defined on a set B, and shows that a small norm implies that the function is amenable to approximation by a neural network with one hidden layer.

Definition 2.1 *For a bounded set $B \subseteq \mathbb{R}^p$ let $\|\omega\|_B = \sup_{x \in B} |\langle \omega, x\rangle|$. For a function $f\colon \mathbb{R}^n \to \mathbb{R}$, define the norm $\|f\|_B^* := \int_{\mathbb{R}^n} \|\omega\|_B \, |\widehat{f}(\omega)| \, d\omega$.*

When $B = B_n$ is the unit ball, $\|\omega\|_B = \|\omega\|_2$. In this case, using Theorem A.3,

$$\|f\|_B^* = \int_{\mathbb{R}^n} \|\omega\| \, |\widehat{f}(\omega)| \, d\omega = \left\| \left\| \omega \widehat{f} \right\|_2 \right\|_1 = \left\| \left\| \widehat{\nabla f} \right\|_2 \right\|_1$$

where for a function $g : \mathbb{R}^n \to \mathbb{R}^n$, $\|g\|_2$ is thought of as a function $\mathbb{R}^n \to \mathbb{R}$, and $\| \|g\|_2 \|_1$ is the L^1 norm of this function.

We would like to define this norm for functions $f\colon B \to \mathbb{R}$. However, the Fourier transform is defined for functions $f\colon \mathbb{R}^n \to \mathbb{R}$. Because we only care about the value of f on B, we allow arbitrary extension outside of B.

Definition 2.2 *Let $B \subseteq \mathbb{R}^n$. Let $\mathcal{F}_B$ be the set of functions for which the Fourier inversion formula holds on B after subtracting out $g(0)$:*[1]

$$\mathcal{F}_B = \left\{ g : \mathbb{R}^n \to \mathbb{R} : \forall x \in B, g(x) = g(0) + \int (e^{i\langle \omega, x\rangle} - 1)\widehat{g}(\omega)\, d\omega \right\}.$$

*Define $\Gamma_B = \{f\colon B \to \mathbb{R} : \exists g, g|_B = f, g \in \mathcal{F}_B\}$, let $\Gamma_B(C)$ be the subset with norm $\le C$ $\Gamma_B(C) = \{f\colon B \to \mathbb{R} : \exists g, g|_B = f, \|g\|_B^* \le C, g \in \mathcal{F}_B\}$. We say that a function $f \in \Gamma_B(C)$ is C-**Barron** on B. For a function $f\colon B \to \mathbb{R}$, let $C_{f,B}$ be the minimal constant for which $f \in \Gamma_{B,C}$:*

$$C_{f,B} := \inf_{g|_B = f, g \in \mathcal{F}_B} \int_{\mathbb{R}^n} \|\omega\|_B \, |\widehat{g}(\omega)| \, d\omega. \tag{6}$$

When the set B is clear, we just write C_f.

This definition is non-algorithmic. How to compute or approximate the Barron constant in general is an open problem. The difficulty stems from the fact that we have to take an infimum over all possible extensions. The Barron constant can be upper-bounded by choosing any extension f, but is more difficult to lower-bound. We will give a technique to lower-bound the Barron constant in Theorem 4.2.

We give some intuition on the Barron constant. First, in order for the Barron constant to be finite, f must be continuously differentiable. Indeed, the inverse Fourier transform of $\omega\widehat{f}(\omega)$ is $-i\nabla f(x)$, and integrability of a function implies continuity of its (inverse) Fourier transform, so ∇f is continuous.

Second, the Barron constant will be larger when $\widehat{f}$ is more "spread out." One can think of $\|g\|_B$ as a kind of L^1 norm. This makes sense in the context of neural networks, because if $f(x) = \sum_{i=1}^k c_i\sigma(\langle a_i, x\rangle + b_i)$ then f has Fourier transform completely supported on the lines in the direction of the a_i.[2] One can think of the Barron constant as a L^1 relaxation of this "sparsity" condition.

Barron's Theorem gives an upper bound on how well a function can be approximated by a neural network with 1 hidden layer of k nodes, in terms of the Barron constant.

For a list of functions with small Barron constant, as well as the effect of various operations on the Barron constant, see (Barron, 1993, §IX). Examples of Barron functions include polynomials of low degree, ridge functions, and linear combinations of Barron functions.

Definition 2.3 *A sigmoidal function is a bounded measurable function $f\colon \mathbb{R} \to \mathbb{R}$ such that* $\lim_{x\to-\infty} f(x) = 0$ *and* $\lim_{x\to\infty} f(x) = 1$.

Theorem 2.4 (Barron, Barron (1993)) *Let $B \subseteq \mathbb{R}^n$ be a bounded set, and μ any probability measure on B. Let $f \in \Gamma_B(C)$ and σ be sigmoidal. There exist $a_i \in \mathbb{R}^n$, $b_i \in \mathbb{R}$, $c_i \in \mathbb{R}$ with $\sum_{i=1}^k |c_i| \le 2C$ such that letting $f_k(x) = \sum_{i=1}^k c_i\sigma(\langle a_i, x\rangle + b_i)$, we have*

$$\|f - f_k\|_\mu^2 := \int_B (f(x) - f_k(x))^2 \, \mu(dx) \le \frac{(2C)^2}{k}.$$

[1] This is a strictly larger set than functions for which the Fourier inversion formula holds.

[2] Here f does not approach 0 as $\|x\| \to \infty$, so the Fourier transform must be understood in the sense of distributions.

Barron's Theorem works for the logistic function (which is sigmoidal), hyperbolic tangent (which is sigmoidal if rescaled to $[0, 1]$), and ReLU up to a factor of 2 in the number of nodes. Even though the ReLU function $\text{ReLU}(x) = \max\{0, x\}$ is not sigmoidal, the linear combination $\text{ReLU}(x) = \text{ReLU}(x) - \text{ReLU}(x-1)$ is.

Note that Barron's Theorem doesn't give approximability tailored to a specific measure μ; it simultaneously gives approximability for *all* μ defined on B, and up to any degree of accuracy. This is why some degree of smoothness is necessary for f: otherwise, μ could be concentrated on the regions where B is not smooth. Note that approximability for all μ will be crucial to the proof of the main theorem (Theorem 3.1). [3]

3. Multilayer Barron's Theorem

3.1. Main theorem

Barron's Theorem says that a Barron function can be approximated by a neural net with 1 hidden layer. From this, it is reasonable to suspect that a composition of l Barron functions can be approximated by a neural network with l hidden layers. Our main theorem says that this is the case; we give a sufficient criterion for a function to be approximated by a neural network with l hidden layers, on any distribution supported in a fixed set K_0.

We note two caveats: first, f_i need to be Lipschitz to prevent the error from blowing up. Second, we will need our functions f_i to be Barron on a slightly expanded set (assumption 3), because an approximation g_i to f_i could take points outside K_i, and we need to control the error for those points.

Given a sequence of functions f_i and $j \ge i$, let $f_{j:i} := f_j \circ f_{j-1} \circ \cdots \circ f_i$.

Theorem 3.1 (Main theorem) *Let $\varepsilon, s > 0$ be parameters, and $l \ge 1$. For $0 \le i \le l$ let $m_i \in \mathbb{N}$. Let $f_i : \mathbb{R}^{m_{i-1}} \to \mathbb{R}^{m_i}$ be functions, μ_0 be any probability distribution on $\mathbb{R}^{m_0}$, and $K_i \subset \mathbb{R}^{m_i}$ be sets.*

Suppose the following hold.

1. *(Support of initial distribution)* $\text{Supp}(\mu_0) \subset K_0$.
2. *(f_i is Lipschitz)* $\text{Lip}(f_i) \le 1$.
3. *(f_i is Barron)* $f_1 \in \Gamma_{K_0}(C_0)$ *and for* $1 \le i \le l$, $f_i \in \Gamma_{K_{i-1}+sB_{m_{i-1}}}(C_i)$.
4. *(f_i takes each set to the next)* $f_i(K_{i-1}) \subseteq K_i$

[3] Although Barron's Theorem seems to require a strong smoothness assumption, we can approximate any continuous function arbitrarily well with a smooth function and then apply Barron's Theorem.
A converse to Barron's Theorem cannot hold in the form stated, because if $\|a_i\|$ is not restricted, then $\sigma(\langle a_i, x\rangle + b_i)$ could have large gradient; the Barron constant of $\phi(\langle a_i, x\rangle + b_i)$ would scale as $\|a_i\|$.
It is natural to ask whether we can choose the a_i to have bounded norm. Barron (Barron, 1993, Theorem 3) shows a version of the theorem that produces a representation with $\|a_i\| \le \tau$, but that incurs an additive error C_τ in the approximation.
Note that the following weak converse holds: the Barron constant of $f = c_0 + \sum_{i=1}^r c_i\sigma(\langle a_i, x\rangle + b_i)$ is bounded by $O(\text{diam}(K) \sum_{i=1}^r |c_i| \|a_i\|)$.

Suppose that the diameter of K_l is D. Then there exists a neural network g with l hidden layers with $\left\lceil \frac{4C_i^2 m_i}{\varepsilon^2} \right\rceil$ nodes on the ith layer, so that

$$\left(\int_{K_0} \|f_{l:1} - g\|^2 \, d\mu_0\right)^{\frac{1}{2}} \le l\varepsilon\sqrt{(2C_l\sqrt{m_l} + D)^2 \frac{l}{3s^2} + 1}. \tag{7}$$

We prove this in Section 3.3. It is crucial to the proof that Barron's Theorem simultaneously gives approximability for *all* probability distributions on a given set.

Note that if K_{l-1} is a ball of radius r, by the way we defined the norm $\|\cdot\|_{K_{l-1}}$ in the Barron constant, C_l will at least scale as $s + r$. If we set s to be on the same order as r, then the RHS of (7) is on the order of $l^{\frac{3}{2}} m_l^{\frac{1}{2}} \varepsilon$.

3.2. Approximating probability distributions

Theorem 3.1 can be interpreted in a very natural way when the aim is to approximate the probability distribution $f_{l:1}(x), x \sim \mu_0$. The Wasserstein distance is a natural distance defined on distributions.

Definition 3.2 *Let μ, ν be two probability distributions on $\mathbb{R}^n$. Let $\Gamma(\mu, \nu)$ denote the set of probability distributions on $\mathbb{R}^n \times \mathbb{R}^n$ whose marginals on the first and second factors are μ and ν respectively. (A distribution $\gamma \sim \Gamma(\mu, \nu)$ is called a **coupling** of μ, ν.) For $1 \le p < \infty$, define the pth **Wasserstein distance** by*

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|_2^p \, d\gamma(x, y)\right)^{\frac{1}{p}}$$

When $p = 1$, this is also known as the "earth mover's distance." One can think of it as the minimum "effort" required to change the distribution of μ to that of ν by shifting probability mass (where "effort" is an integral of mass times distance).

Corollary 3.3 *Keep the notation in Theorem 3.1 and suppose the diameter of the set $f_{l:1}(K_0)$ is D. Then the Wasserstein distance between the distribution $f_{l:1}(X)(X \sim \mu_0)$ and $g(X), (X \sim \mu_0)$ is at most $l\varepsilon\sqrt{1 + (2C_l\sqrt{m_l} + D)^2 \frac{l}{3s^2}}$.*

The proof of this is simple: observe that $(f_{l:1}(X), g(X))$, $X \sim \mu_0$ defines a coupling between the distributions. Thus by Theorem 3.1 the W_2 Wasserstein distance is at most

$$\left[\mathop{\mathbb{E}}_{X \sim \mu_0} \|f_{l:1}(X) - g(X)\|^2\right]^{\frac{1}{2}} \le l\varepsilon\sqrt{(2C_l\sqrt{m_l} + D)^2 \frac{l}{3s^2} + 1}.$$

The Wasserstein distance is a suitable metric in the context of GANs (Arjovsky and Bottou (2017), Arjovsky et al. (2017)). One way to model a discriminator is as a function f in a certain class F that maximizes the difference between $\mathbb{E}f$ on the real distribution μ and the generated distribution ν,

$$\sup_{f \in F} \left| \mathop{\mathbb{E}}_{x \sim \mu} f(x) - \mathop{\mathbb{E}}_{y \sim \nu} f(y) \right|. \tag{8}$$

This is called the maximal mean discrepancy (Kifer et al. (2004), Dziugaite et al. (2015)). The Wasserstein distance captures the idea that if two distributions are close, then it is hard for such a Lipschitz discriminator to tell the difference, as the following lemma shows.

Lemma 3.4 (Properties of Wasserstein metric) *For any two distributions μ, ν over $\mathbb{R}^n$, $W_1(\mu,\nu) \le W_2(\mu,\nu)$. Moreover, for any Lipschitz function $f\colon \mathbb{R}^n \to \mathbb{R}$,*

$$\left| \mathop{\mathbb{E}}_{x\sim\mu} f(x) - \mathop{\mathbb{E}}_{y\sim\nu} f(y) \right| \le \operatorname{Lip}(f) W_1(\mu,\nu). \tag{9}$$

Proof is deferred to Appendix C. In the context of Corollary 3.3, Lemma 3.4 says that the distribution generated by $f_{l:1}$ and by the neural network cannot be distinguished by a Lipschitz function. Arjovsky et al. (2017) discuss why the class of Lipschitz functions is a good choice in comparison to other classes. For instance, if we maximize over the class of indicator functions (of measurable sets) instead, (8) becomes the total variation (TV) distance, which is unstable under perturbations to the function generating the distribution. In particular, the TV distance is discontinuous under perturbations of distributions supported on lower-dimensional subsets of the ambient space $\mathbb{R}^n$.

3.3. Proof of main theorem

To prove Theorem 3.1 we first prove the following theorem.

Theorem 3.5 *Keep conditions 1–4 and the notation of Theorem 3.1. Then there exists a neural network g with l hidden layers and $S \subset \mathbb{R}^{m_0}$ satisfying $\mu_0(S) \ge 1 - \left(\sum_{i=1}^{l-1} i^2\right)\frac{\varepsilon^2}{s^2}$ so that*

$$\left(\int \mathbb{1}_S \left\|f_{l:1} - g\right\|^2 \, d\mu_0\right)^{\frac{1}{2}} \le l\varepsilon \tag{10}$$

Proof Let $r_i = \left\lceil \frac{4C_i^2 m_i}{\varepsilon^2} \right\rceil$. We will show that we can take $g = g_{l:1}$, where $g_1, \ldots, g_l$ are functions defined by

$$g_i : \mathbb{R}^{m_{i-1}} \to \mathbb{R}^{m_i} \tag{11}$$

$$(g_i(x))_j = c_{ij0} + \sum_{k=1}^{r_i} c_{ijk}\sigma(\langle a_{ijk}, x\rangle + b_{ijk}), \tag{12}$$

for some parameters $c_{ijk}, b_{ijk} \in \mathbb{R}$, $a_{ijk} \in \mathbb{R}^{m_{i-1}}$. Note that each g_i is a neural net with one hidden layer and a linear output layer. When the next layer g_{i+1} is applied to the output y of g_i, first linear functions $\langle a_{i+1,j,k}, y\rangle + b_{i+1,j,k}$ are applied; these linear functions can be collapsed with the linear output layer of g_i. Thus only one hidden layer is added each time.

We prove the statement by induction on l. For $l = 1$, the theorem follows directly from Barron's Theorem 2.4, using assumptions 1 and 3.

For the induction step, assume we have functions $g_1, \ldots, g_{l-1}$ satisfying the conclusion for $f_1, \ldots, f_{l-1}$. Let S_{l-1} be the set in the conclusion. Apply Barron's Theorem 2.4 to f_l to get that that for each $1 \le j \le m_l$, for any μ supported on a set $K'_{l-1} \subseteq \mathbb{R}^{m_{l-1}}$ and any $r_l \in \mathbb{N}$, there exists a neural net $g_{l,j}$ with 1 hidden layer with r_l nodes such that

$$\left(\int_{\mathbb{R}^{m_{l-1}}} \left[(f_l)_j - (g_l)_j\right]^2 \, d\mu\right)^{\frac{1}{2}} \le \frac{2C_{f_l,K'_{l-1}}}{\sqrt{r_l}}.$$

Note it is vital here that Barron's Theorem applies to *any* distribution μ supported on K'_{l-1}. Let $S_l = S_{l-1} \cap \left\{x : g_{l-1:1}(x) \in K_{l-1} + sB_{m_{l-1}}\right\}$. Apply Barron's Theorem with $K'_l = K_l + sB_{m_l}$,

$r_l = \left\lceil \frac{4C_l^2 m_l}{\varepsilon^2} \right\rceil$. $\mu = g_{l-1:1*}(\mathbb{1}_{S_l}\mu_0)$. [4] We have that μ is supported on $g_{l-1:1}(S_l) \subseteq K_{l-1} + sB_{m_{l-1}} = K'_{l-1}$, as required, and f_l is C_l-Barron on this set by assumption 3. (Note that μ is not a probability measure because it was restricted to the set $g_{l-1:1}(S_l)$, but it is a nonnegative measure with total L^1 mass at most 1. Because Barron's Theorem holds for any probability measure, it also holds for these measures.) The conclusion of Barron's Theorem gives $(g_l)_j$ such that

$$\left(\int_{\mathbb{R}^{m_{l-1}}} [(f_l)_j - (g_l)_j]^2 \, d(g_{l-1:1*}(\mathbb{1}_{S_l}\mu_0))\right)^{\frac{1}{2}} \le \frac{2C_l}{\sqrt{r_l}} \le \frac{\varepsilon}{\sqrt{m_l}} \tag{13}$$

$$\implies \left(\int_{\mathbb{R}^{m_{l-1}}} \|f_l - g_l\|^2 \, d(g_{l-1:1*}(\mathbb{1}_{S_l}\mu_0))\right)^{\frac{1}{2}} \le \varepsilon \tag{14}$$

We bound by the triangle inequality

$$\begin{aligned}
&\left(\int_{\mathbb{R}^m} \mathbb{1}_{S_l} \|f_{l:1} - g_{l:1}\|^2 \, d\mu_0\right)^{\frac{1}{2}} \\
&\le \left(\int_{\mathbb{R}^m} \mathbb{1}_{S_l} \|f_l \circ f_{l-1:1} - f_l \circ g_{l-1:1}\|^2 \, d\mu_0\right)^{\frac{1}{2}} + \left(\int_{\mathbb{R}^m} \mathbb{1}_{S_l} \|f_l \circ g_{l-1:1} - g_l \circ g_{l-1:1}\|^2 \, d\mu_0\right)^{\frac{1}{2}} \\
&\le \left(\int_{\mathbb{R}^m} \mathbb{1}_{S_l} \|f_l \circ f_{l-1:1} - f_l \circ g_{l-1:1}\|^2 \, d\mu_0\right)^{\frac{1}{2}} + \left(\int_{\mathbb{R}^{m_{l-1}}} \|f_l - g_l\|^2 \, dg_{l-1:1*}(\mathbb{1}_{S_l}\mu_0)\right)^{\frac{1}{2}} \\
&\le \mathrm{Lip}(f_l)\left(\int_{\mathbb{R}^m} \mathbb{1}_{S_l} \|(f_{l-1:1} - g_{l-1:1})\|^2 \, d\mu_0\right)^{\frac{1}{2}} + \varepsilon \\
&\le \mathrm{Lip}(f_l)\left(\int_{\mathbb{R}^m} \mathbb{1}_{S_{l-1}} \|(f_{l-1:1} - g_{l-1:1})\|^2 \, d\mu_0\right)^{\frac{1}{2}} + \varepsilon \\
&\le 1 \cdot (l-1)\varepsilon + \varepsilon = l\varepsilon
\end{aligned}$$

The last inequality holds by assumption 2 and the induction hypothesis.

To finish, we have to check that $\mu_0(S_l) \ge 1 - \left(\sum_{i=1}^{l-1} i^2\right)\frac{\varepsilon^2}{s^2}$. As above, we have that

$$\int \mathbb{1}_{S_{l-1}} \|f_{l-1:1} - g_{l-1:1}\|^2 \, d\mu_0 \le (l-1)^2\varepsilon^2$$

by the induction hypothesis. Also, $f_{l-1:1}(x) \in K_{l-1}$ for all $x \in \mathrm{Supp}(\mu_0)$ by assumption 4. Thus by Markov's inequality and the induction hypothesis on S_{l-1},

$$\begin{aligned}
&\mu_0(S_{l-1} \cap \left\{x : x \notin K_{l-1} + sg_{l-1:1}(B_{m_{l-1}})\right\}) \\
&\le \mu_0(S_{l-1} \cap \{x : \|f_{l-1:1}(x) - g_{l-1:1}(x)\| \ge s\}) \le \frac{(l-1)^2\varepsilon^2}{s^2}
\end{aligned}$$

Therefore $\mu_0(S_l) \le \mu_0(S_{l-1}) - \frac{(l-1)^2\varepsilon^2}{s^2} \le 1 - \left(\sum_{i=1}^{l-1} i^2\right)\frac{\varepsilon^2}{s^2}$. ■

It is inelegant to have to exclude the sets S_l. The main theorem is a statement that doesn't involve the sets S_l. We achieve this by using the trivial bound on S_l^c.

[4] The pushforward of a measure μ by a function f is denoted by $f_*\mu$ and defined by $f_*\mu(S) = \mu(f^{-1}(S))$. Here, $g_{l-1:1*}(\mathbb{1}_{S_l}\mu_0)(S) = \mu_0(g_{l-1:1}^{-1}(S) \cap S_l)$.

Proof [Proof of Theorem 3.1] The functions $g_1, \dots, g_l$ in Theorem 3.5 satisfy $\int_{S_l} \|f_{l:1} - g_{l:1}\|^2 \, d\mu_0 \le l^2\varepsilon^2$. The range of $g_l = ((g_l)_1, \dots, (g_l)_{m_l})$ is contained in a set of diameter $2C_l\sqrt{m_l}$ because the function σ has range contained in $[0,1]$ and Barron's Theorem gives functions $(g_l)_j$, $1 \le j \le m_l$, with $\sum_{k=1}^{r} |c_{ljk}| \le 2C_l$.

Choose a constant vector k to minimize $\int_{S_l} \|f_{l:1}(x) - g_{l:1}(x) - k\|^2 \, d\mu_0$ and replace g_l with $g_l + k$. Note that now, the range of g_l and f_l necessarily overlap; otherwise a further translation will decrease this error. We still have $\int_{S_l} \|f_{l:1} - g_{l:1}\|^2 \, d\mu_0 \le l^2\varepsilon^2$. Moreover, $\|g_l(x) - f_l(x)\| \le 2C_l\sqrt{m_l} + D$ for any $x \in K_0$.

Now we have (using $\mu_0(S_l^c) \le \left(\sum_{i=1}^{l-1} i^2\right) \frac{\varepsilon^2}{s^2} \le \frac{l^3\varepsilon^2}{3s^2}$)

$$\int_{K_0} \|f_{l:1} - g_{l:1}\|^2 \, d\mu_0 \le \int_{S_l} \|f_{l:1} - g_{l:1}\|^2 \, d\mu_0 + \int_{S_l^c} \|f_{l:1} - g_{l:1}\|^2 \, d\mu_0 \tag{15}$$

$$\le l^2\varepsilon^2 + (2C_l\sqrt{m_l} + D)^2 \frac{l^3\varepsilon^2}{3s^2}. \tag{16}$$

Taking square roots gives the theorem. ■

4. Separation between Barron functions and composition of Barron functions

In this section we produce an explicit function $f \colon \mathbb{R}^n \to \mathbb{R}$ that is a composition of two $\mathrm{poly}(n)$-Barron functions, but is not $O(c^n)$-Barron for some $c > 1$.

Theorem 4.1 *For any $n \equiv 3 \pmod 4$ and $c > 1$, there exists a function f and $C_2 > 0$ such that*

1. *(f is not Barron) $C_{f, C_2 n B_n} \ge c^n$.*
2. *(f is the composition of 2 Barron functions) $f = j \circ k$ where for all $r, s > 0$, $k : \mathbb{R}^n \to \mathbb{R}$ is $O(nr^3)$-Barron on rB_n, and $j : \mathbb{R} \to \mathbb{R}$ is $O(sn^2)$-Barron on sB_1.*

The condition $n \equiv 3 \pmod 4$ is not necessary; we include it only to avoid case analysis.

Note that this theorem gives a separation between Barron functions and compositions of Barron functions, and does not give a separation between distributions expressible by Barron functions and compositions of Barron functions. The analogous question for distributions is an open problem.

We will choose f to be a certain radial function $f = f_1(\|x\|)$ defined in Section 4.1.[5] In order for f to have large Barron constant, it is necessary for $\int_{\mathbb{R}^n} \|\omega\|_2 |\widehat{f}(\omega)| \, d\omega$ to be large, i.e. for $\widehat{f}$ to have significant mass far away from the origin. We ensure this holds by choosing f to change sharply in the radial direction. This means $\widehat{f}$ has mass far away from the origin. Moreover, $\widehat{f}$ is radial because f is radial, so $\widehat{f}$ has significant mass in a large shell.

However, lower-bounding $\int_{\mathbb{R}^n} \|\omega\|_2 |\widehat{f}(\omega)| \, d\omega$ is not sufficient because the definition of the Barron constant requires us to bound this quantity over all extensions of f.

To solve this problem, we give a technique to lower bound the Barron constant in Section 4.2 (Theorem 4.2). Although we cannot certify f is Barron by showing $\int_{\mathbb{R}^n} \left\| \widehat{\nabla f}(\omega) \right\| \, d\omega = \int_{\mathbb{R}^n} \|\omega\|_2 |\widehat{f}(\omega)| \, d\omega$ is large, it suffices to show $\int_{\mathbb{R}^n} \left\| \widehat{(\nabla f) g}(\omega) \right\| \, d\omega$ is large for a judiciously chosen g. We use this to show that f is not Barron in Section E.1 (Theorem E.4).

[5] For any radial function $a : \mathbb{R}^n \to \mathbb{R}$, we write $a_1 : \mathbb{R} \to \mathbb{R}$ for the function such that $a(x) = a_1(\|x\|)$.

We will see in Section 4.3 (Theorem 4.4) that f is a composition of two Barron functions $x \mapsto \|x\|^2$ and $y \mapsto f_1(\sqrt{y})$. The function $x \mapsto \|x\|^2$ is Barron because it is a polynomial. The function $y \mapsto f_1(\sqrt{y})$ is a function in 1 variable, and it is much easier for a 1-dimensional function h to be Barron as bounds on h, h', and h'' suffice (Lemma A.6).

Our result is similar to the construction in Eldan and Shamir (2015) of an explicit function that can be approximated by a 3-layer neural net but cannot be approximated (to better than constant error) by any 2-layer neural net with subexponential number of units. Eldan and Shamir (2015) use a different Fourier criterion in order to prove a certain function is not computable by a two-layer neural network.

Roughly speaking, Eldan and Shamir implicitly show that for a specific probability measure that they chose (φ^2, where $\widehat{\varphi} = \mathbb{1}_{R_n B_n}$, where R_n is chosen so that $\text{Vol}(R_n B_n) = 1$), a necessary criterion for f to be approximated by a 2-layer neural network with k nodes is that most of its mass is concentrated in k "tubes" $\bigcup_{i=1}^k (\text{span}\{v_i\} + R_n B_n)$. (See (Eldan and Shamir, 2015, Proposition 13, Claim 15, Lemma 16).) The idea can be adapted to other measures. The main difference from Barron's Theorem is that their criterion is a necessary condition for approximability (so useful to show lower bounds), is measure-specific (rather than agnostic to the measure), and is more similar to a "sparsity" condition than a "L^1 measure" as in Barron's Theorem.

4.1. Definition of f

Let $f_1 : \mathbb{R} \to \mathbb{R}$ be a function such that f_1 is nonnegative, $\text{Supp}(f_1) \subseteq [K_1, K_1+\varepsilon]$, $\int_0^\infty f_1(x)\,dx = 1$, and $|f_1^{(i)}| = O\left(\frac{1}{\varepsilon^{i+1}}\right)$ for all $i = 0, 1, 2$. This function exists by Lemma D.1(1). We will choose K_1, ε depending on n.

By Theorem A.5,

$$\widehat{f}(\omega) = \frac{1}{2\pi}\left(\frac{1}{2\pi\,\|\omega\|}\right)^{\frac{n}{2}-1} \int_0^\infty r^{\frac{n}{2}-1} f_1(r) J_{\frac{n}{2}-1}(\|\omega\|\,r)\,dr. \tag{17}$$

We will choose $[K_1, K_1 + \varepsilon]$ to be an interval on which $J_{\frac{n}{2}}(\|\omega\|\,r)$ is large and positive for some large $\|\omega\|$.

We use the notation of Lemma B.1. For $x \geq n$,

$$(f_{n,x}x)' = \frac{x}{\sqrt{x^2 - \left(\frac{n^2-1}{4}\right)}} - \frac{\sqrt{n^2-1}}{2} \cdot \frac{1}{\sqrt{1 - \frac{n^2-1}{4x^2}}} \cdot \frac{-\sqrt{n^2-1}}{2x^2} = \sqrt{1 - \frac{n^2-1}{4x^2}} \in \left[\sqrt{\frac{3}{4}}, 1\right].$$

Let $K_3 = C_3\sqrt{n}$ for some C_3 to be chosen. In every interval of length $\geq \frac{4\pi}{K_3\sqrt{3/4}}$ there is an interval of length $\geq \frac{\pi}{K_3}$ on which

$$\cos\left(-\frac{(n+1)\pi}{4} + f_{d,K_3 r}K_3 r\right) \geq \frac{1}{\sqrt{2}}. \tag{18}$$

Let $[K_1, K_1 + \varepsilon]$ be the first such interval with $K_1 \geq C_1\sqrt{n}$, where C_1 is a constant to be chosen. Note we have $K_1 \sim C_1\sqrt{n}$ and $\varepsilon = \Theta\left(\frac{1}{K_3}\right)$.

4.2. A technique to lower bound the Barron constant

The main difficulty in showing a function is not Barron is to lower bound the integral

$$\int_{\mathbb{R}^n} \|\omega\| \, |\widehat{F}(\omega)| \, d\omega = \int_{\mathbb{R}^n} \left\| \widehat{\nabla F}(\omega) \right\| \, d\omega$$

over *all* extensions F of f. In general, it is not known how to calculate the infimum over all extensions.

Theorem 4.2 gives us a way to lower-bound the Barron constant for f over a ball rB_n. The idea is the following. Instead of bounding $\int_{\mathbb{R}^n} \left\| \widehat{\nabla F}(\omega) \right\| \, d\omega$ for every extension F, we choose g with support in B and compute $\int_{\mathbb{R}^n} \left\| \widehat{(\nabla F)g}(\omega) \right\| \, d\omega$. This does not depend on the extension F because $(\nabla F)g = (\nabla f)g$. It turns out that we can bound $\int_{\mathbb{R}^n} \left\| \widehat{\nabla F}(\omega) \right\| \, d\omega$ in terms of $\int_{\mathbb{R}^n} \left\| \widehat{(\nabla F)g}(\omega) \right\| \, d\omega$.

Theorem 4.2 *If f is differentiable, then for any g such that* $\mathrm{Supp}(g) \subseteq rB_n$ *and* $g, \widehat{g} \in L^1(\mathbb{R}^n)$,

$$C_{f,rB_n} \geq r \frac{\int_{\mathbb{R}^n} |\widehat{(\nabla f)g}(\omega)| \, d\omega}{\int_{\mathbb{R}^n} |\widehat{g}(\omega)| \, d\omega}$$

Note that g is a function that we are free to choose. To use the theorem we will choose g with $\mathrm{Supp}(g) \subseteq C_2 n B_n$ and $\int_{\mathbb{R}^n} |\widehat{g}(\omega)| \, d\omega$ small. This theorem is similar to (Barron, 1993, §IX.11), which bounds the Barron constant of a product of two functions. We defer the proof to Appendix E.

To use this bound for a function f, we need to judiciously choose the function g. Let b be the "bump" function given by Lemma D.1(3) for $m = \frac{n+1}{2}$. This function has the properties that $b(x) = 1$ for $x \in [-1,1]$, $b(x) = 0$ for $|x| \geq 2$, and for $k \leq m$, $b^{(k)}(x) \leq (n+1)^k$. Let $g_1(x) = b_{(K_2)}(x) = b\left(\frac{x}{K_2}\right)$ and $g(x) = g_1(\|x\|)$ for $K_2 = C_2 n$, where C_2 is a constant to be chosen.

In Appendix E, we show the following lemma that bounds the Barron constant for f.

Lemma 4.3 *For $n \equiv 3 \pmod 4$ and constants C_1, C_2, C_3 such that $C_1 C_3 \geq \frac{3}{2}$, $C_2 > C_1 \geq 1$, $C_3 \geq 1$, the functions f, g we choose satisfy*

$$\int_{\mathbb{R}^n} |\widehat{g}(\omega)| \, d\omega = O((5eC_2)^{\frac{n}{2}}), \tag{19}$$

$$\int_{\mathbb{R}^n} \left\| \widehat{(\nabla f)g}(\omega) \right\| \, d\omega = \Omega(C_1^{\frac{n}{2}-3} C_3^{\frac{n}{2}} n^{-\frac{1}{2}} e^{\frac{n}{2}}). \tag{20}$$

As a result the Barron constant $C_{f,2K_2B_n} \geq \Omega\left(2^{-n} C_1^{\frac{n}{2}-3} C_3^{\frac{n}{2}} C_2^{-\left(\frac{n}{2}-1\right)} n^{\frac{1}{2}}\right)$.

Therefore, as long as we choose C_3 to be large enough this constant is exponentially large. The constraint that $n \equiv 3 \pmod 4$ is only there to avoid case analysis. We give the proof in Section E.

4.3. h is a composition of Barron functions

We can write f as the composition of a function that computes the square norm, and a one dimensional function. The Barron constant for both functions can be bounded by polynomials.

Lemma 4.4 *Suppose that $C_1 < C_3$. f is the composition of the two functions*

$$x \mapsto \|x\|^2 \qquad \mathbb{R}^n \to \mathbb{R} \tag{21}$$

$$y \mapsto f_1(\sqrt{y}) \qquad \mathbb{R} \to \mathbb{R}. \tag{22}$$

The function $x \mapsto \|x\|^2$ satisfies $C_{\|x\|^2, rB_n} \leq O(nr^3)$ and the function $y \mapsto f_1(\sqrt{y})$ satisfies $C_{f_1(\sqrt{y}),[-s,s]} = O(sC_1^{\frac{1}{2}} C_3^{\frac{3}{2}} n^2)$ for any s.

Intuitively, the proof uses the fact that polynomials are Barron, and all "nice" one dimensional functions are Barron. We leave the detailed proofs in Section E. Now it is easy to see the separation:
Proof [of Theorem 4.1] By Lemma 4.3, we know we can choose C_3 large enough so that the Barron constant for f is exponential. On the other hand, by Lemma 4.4 we know f is a composition of two Barron functions. ■

5. Conclusion

In this paper we show if a generative model can be expressed as the composition of n Barron functions, then it can be approximated by a $n+1$-layer neural network. Along the way we proved a multi-layer version of the Barron's Theorem (Barron, 1993), and a key observation is to use Wasserstein distance W^2 as the distance measure between distributions. This partly explains the expressive power of neural networks as generative models. However, there are still many open problems: what natural transformations can be represented by a composition of Barron functions? Is there a separation between composition of n Barron functions and composition of $n+1$ Barron functions? How can we learn such a representation efficiently? We hope this paper serves as a first step towards understanding the power of deep generative models.

Acknowledgement

This work was done in part while the authors were visiting the Simons Institute for the Theory of Computing at UC Berkeley.

References

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *NIPS 2016 Workshop on Adversarial Training. In review for ICLR*, volume 2016, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993. ISSN 00189448. doi: 10.1109/18.256500.

Andrew R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14(1):115–133, 1994. ISSN 08856125. doi: 10.1007/BF00993164.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Nadav Cohen, Or Sharir, and Amnon Shashua. On the expressive power of deep learning: A tensor analysis. *arXiv preprint arXiv:1509.05009*, 554, 2015.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, 2(4):303–314, 1989.

Amit Daniely. Depth separation for neural networks. *arXiv preprint arXiv:1702.08489*, 2017.

Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*, 2015.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. *arXiv preprint arXiv:1512.03965*, 2015.

Ken-Ichi Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural networks*, 2(3):183–192, 1989.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *CoRR abs/1506.08473*, 2015.

Daniel M Kane and Ryan Williams. Super-linear gate and super-quadratic wire lower bounds for depth-two and depth-three threshold circuits. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pages 633–643. ACM, 2016.

Leonid Vasilevich Kantorovich and G Sh Rubinstein. On a space of completely additive functions. *Vestnik Leningrad. Univ*, 13(7):52–59, 1958.

Michael Kearns, Yishay Mansour, Dana Ron, Ronitt Rubinfeld, Robert E Schapire, and Linda Sellie. On the learnability of discrete distributions. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 273–282. ACM, 1994.

Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 180–191. VLDB Endowment, 2004.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Ilia Krasikov. Approximations for the bessel and airy functions with an explicit error term. *LMS Journal of Computation and Mathematics*, 17(01):209–225, 2014.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015. doi: 10.1016/j.neunet.2014.09.003. Published online 2014; based on TR arXiv:1404.7828 [cs.NE].

Matus Telgarsky. Benefits of depth in neural networks. *arXiv preprint arXiv:1602.04485*, 2016.

Appendix A. Background from Fourier Analysis

The Fourier transform is defined in (5).

Theorem A.1 (Fourier inversion) *For continuous f such that $f \in L^1(\mathbb{R}^n)$ and $\widehat{f} \in L^1(\mathbb{R}^n)$,*

$$f(x) = \int \widehat{f}(x) e^{i\langle \omega, x\rangle}\, dx. = (2\pi)^n \hat{\hat{f}}^{\vee}$$

Theorem A.2 (Plancherel's Theorem) *For $f, g : \mathbb{R}^n \to \mathbb{C}$ such that $f, g \in L^1(\mathbb{R}^n) \cap L^2(\mathbb{R}^n)$,*

$$\int_{\mathbb{R}^n} f(x)\overline{g(x)}\, dx = \int_{\mathbb{R}^n} (2\pi)^n \widehat{f}(\omega)\overline{\widehat{g}(\omega)}\, d\omega.$$

Theorem A.3 (Fourier transform of derivative) *For differentiable $f : \mathbb{R}^n \to \mathbb{R}$, $f \in L^1(\mathbb{R}^n)$,*

$$\widehat{\nabla f}(x) = ix\widehat{f}(x).$$

For $f : \mathbb{R}^n \to \mathbb{R}$ such that $f, \|x\| f \in L^1(\mathbb{R}^n)$,

$$(xf)^{\wedge} = i\nabla \widehat{f}(x).$$

Theorem A.4 (Fourier transform of convolution) *For $f, g \in L^1(\mathbb{R}^n)$*

$$\widehat{f * g}(x) = \widehat{f}(\omega)\widehat{g}(\omega) \tag{23}$$

For $f, g \in L^1(\mathbb{R}^n)$ with $fg, \widehat{f}, \widehat{g} \in L^1(\mathbb{R}^n)$,

$$\widehat{fg}(x) = (\widehat{f} * \widehat{g})(\omega). \tag{24}$$

Theorem A.5 (Fourier transform of radial function) *Suppose $f(x) = f_1(\|x\|)$ where $f \in L^1(\mathbb{R}^n)$, $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$. Then*

$$\widehat{f}(\omega) = \frac{1}{2\pi}\left(\frac{1}{2\pi\|\omega\|}\right)^{\frac{n}{2}-1} \int_0^\infty r^{\frac{n}{2}-1} f_1(r) J_{\frac{n}{2}-1}(\|\omega\| r)\, dr.$$

where J_α is the Bessel function of order α.

Lemma A.6 (L^1 bound on Fourier transform)

1. *Let $k \geq \frac{n+1}{2}$ and k be even. Then for $g : \mathbb{R}^n \to \mathbb{R}$ that is k times differentiable,*

$$\int_{\mathbb{R}^n} \|\widehat{g}(\omega)\|\, d\omega \leq \left(\frac{\Gamma\left(\frac{1}{2}\right)}{2^n \pi^{\frac{n}{2}} \Gamma\left(\frac{n+1}{2}\right)}\right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^n} [(I - \Delta)^{\frac{k}{2}} g(x)]^2\, dx\right)^{\frac{1}{2}}. \tag{25}$$

2. *Let $h : \mathbb{R} \to \mathbb{R}$ be once or twice differentiable, respectively. Then*

$$\int_{-\infty}^{\infty} |\widehat{h}(\omega)|\, d\omega \leq 2^{-\frac{1}{2}} \left(\int_{-\infty}^{\infty} |h|^2 + |h'|^2\, dx\right)^{\frac{1}{2}} \tag{26}$$

$$\int_{-\infty}^{\infty} |\omega\widehat{h}(\omega)|\, d\omega \leq 2^{-\frac{1}{2}} \left(\int_{-\infty}^{\infty} |h'|^2 + |h''|^2\, dx\right)^{\frac{1}{2}}. \tag{27}$$

Proof By Cauchy-Schwarz and the fact that $\int_{\mathbb{R}^n} \frac{1}{(1+\|\omega\|^2)^{\frac{n+1}{2}}}\, d\omega = \frac{\pi^{\frac{n}{2}}\Gamma(\frac{1}{2})}{\Gamma(\frac{n+1}{2})}$ (this is used e.g. to define the Cauchy probability distribution)

$$\int_{\mathbb{R}^n} \|\widehat{g}(\omega)\|\, d\omega \le \left(\int_{\mathbb{R}^n} \frac{1}{\left(1+\|\omega\|^2\right)^k}\, d\omega\right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^n} (1+\|\omega\|^2)^k |\widehat{g}(\omega)|^2\, d\omega\right)^{\frac{1}{2}} \tag{28}$$

$$\le \left(\int_{\mathbb{R}^n} \frac{1}{\left(1+\|\omega\|^2\right)^{\frac{n+1}{2}}}\, d\omega\right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^n} (1+\|\omega\|^2)^k |\widehat{g}(\omega)|^2\, d\omega\right)^{\frac{1}{2}} \tag{29}$$

$$\le \left(\frac{\pi^{\frac{n}{2}}\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}\right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^n} \left|(1+\|\omega\|^2)^{\frac{k}{2}}\widehat{g}(\omega)\right|^2\, d\omega\right)^{\frac{1}{2}} \tag{30}$$

$$\le \left(\frac{\pi^{\frac{n}{2}}\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}\right)^{\frac{1}{2}} (2\pi)^{-\frac{n}{2}} \left(\int_{\mathbb{R}^n} [(I-\Delta)^{\frac{k}{2}} g(x)]^2\, dx\right)^{\frac{1}{2}} \tag{31}$$

where in the last step we used Theorem A.2 and the calculation

$$\widehat{\Delta g} = \left(\sum_{i=1}^n \frac{\partial^2}{\partial {x_i}^2} g\right)^{\wedge} = -\sum_{i=1}^n \omega_i^2 \widehat{g}(\omega) = -\|\omega\|^2\, \widehat{g}(\omega).$$

For the second part, again by Cauchy-Schwarz and $\widehat{h'}(\omega) = i\omega h(\omega)$,

$$\int_{-\infty}^{\infty} |\widehat{h}(\omega)|\, d\omega \le \left(\int_{-\infty}^{\infty} \frac{1}{1+|\omega|^2}\, d\omega \int_{-\infty}^{\infty} |\widehat{h}(\omega)|^2 (1+|\omega|^2)\, d\omega\right)^{\frac{1}{2}} \tag{32}$$

$$\le \sqrt{\pi}\left(\int_{-\infty}^{\infty} |\widehat{h}|^2 + |\widehat{h'}|^2\, d\omega\right)^{\frac{1}{2}} \tag{33}$$

$$\le \sqrt{\pi}(2\pi)^{-\frac{1}{2}} \left(\int_{-\infty}^{\infty} |h|^2 + |h'|^2\, dx\right)^{\frac{1}{2}}. \tag{34}$$

This gives the first equation. To get the second, replace h with h'. ■

Appendix B. Bessel functions

We will need some facts about Bessel functions $J_\alpha(x)$, $\alpha \in \mathbb{R}$. $J_\alpha(x)$ has an oscillating shape like a damped sinusoid.

Lemma B.1 ((Krasikov, 2014, Theorem 5), (Eldan and Shamir, 2015, Lemma 21)) *If $d \ge 2$ and $x \ge d$, then*

$$\left| J_{d/2}(x) - \sqrt{\frac{2}{\pi c_{d,x} x}} \cos\left(-\frac{(d+1)\pi}{4} + f_{d,x} x\right)\right| \le x^{-3/2},$$

where

$$c_{d,x} = \sqrt{1 - \frac{d^2-1}{4x^2}} \quad , \quad f_{d,x} = c_{d,x} + \frac{\sqrt{d^2-1}}{2x} \arcsin\left(\frac{\sqrt{d^2-1}}{2x}\right).$$

Moreover, assuming $x \geq d$*,*

$$1 \geq c_{d,x} \geq 1 - \frac{0.15\,d}{x} \geq 0.85$$

and

$$1.3 \geq 1 + \frac{0.3\,d}{x} \geq f_{d,x} \geq 1 - \frac{0.15\,d}{x} \geq 0.85.$$

Lemma B.2 ((Eldan and Shamir, 2015, Lemma 20)) *For any* $\alpha \geq 1$ *and* $x \geq 3\alpha$*,* $J_\alpha(x)$ *is 1-Lipschitz in* x*.*

Appendix C. Properties of Wasserstein Distance

Lemma C.1 (Lemma 3.4 restated) *For any two distributions* μ, ν *over* $\mathbb{R}^n$*,*

$$W_1(\mu,\nu) \leq W_2(\mu,\nu). \tag{35}$$

Moreover, for any Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$*,*

$$\left| \mathop{\mathbb{E}}_{x\sim\mu} f(x) - \mathop{\mathbb{E}}_{y\sim\nu} f(y) \right| \leq \operatorname{Lip}(f) W_1(\mu,\nu). \tag{36}$$

Proof Let $\gamma \in \Gamma(\mu,\nu)$ be a coupling of μ, ν. Then by the Cauchy-Schwarz inequality,

$$\begin{aligned} W_1(\mu,\nu) &\leq \int_{\mathbb{R}^n\times\mathbb{R}^n} \|x-y\|_2 \, d\gamma(x,y) && (37)\\ &\leq \left(\int_{\mathbb{R}^n\times\mathbb{R}^n} \|x-y\|_2^2 \, d\gamma(x,y)\right)^{\frac{1}{2}} \underbrace{\left(\int_{\mathbb{R}^n\times\mathbb{R}^n} d\gamma\right)^2}_{1}. && (38)\end{aligned}$$

The infimum of (38) over all couplings $\gamma \sim \Gamma(\mu,\nu)$ is exactly $W_2(\mu,\nu)$. This shows (35).

Now for any $\gamma \in \Gamma(\mu,\nu)$, because its marginals are μ and ν,

$$\begin{aligned} \left| \mathop{\mathbb{E}}_{x\sim\mu} f(x) - \mathop{\mathbb{E}}_{y\sim\nu} f(y) \right| &= \left| \int_{\mathbb{R}^n\times\mathbb{R}^n} f(x) - f(y) \, d\gamma(x,y) \right| && (39)\\ &\leq \operatorname{Lip}(f) \int_{\mathbb{R}^n\times\mathbb{R}^n} \|f(x) - f(y)\|_2 \, d\gamma(x,y). && (40)\end{aligned}$$

The Lipschitz constant is with respect to the L^2 norm because we use the L^2 norm to measure the distance between $f(x)$ and $f(y)$. Taking the infimum of (40) gives (36). ■

In fact, (36) is sharp when μ, ν have bounded support. The duality theorem of Kantorovich and Rubinstein (Kantorovich and Rubinstein, 1958) says that

$$W_1(\mu,\nu) = \sup\left\{ \mathop{\mathbb{E}}_{x\sim\mu} f(x) - \mathop{\mathbb{E}}_{y\sim\nu} f(y) : f : \mathbb{R}^n \to \mathbb{R}, \operatorname{Lip}(f) \leq 1 \right\}.$$

Appendix D. Test functions

For a function f, let $f_{(K)}(x) := f\left(\frac{x}{K}\right)$.

Lemma D.1 *Let $m \geq 2$ be a given positive integer.*

1. *There exists a function $g : \mathbb{R} \to \mathbb{R}$ with the following properties.*
   - *(a)* $g \geq 0$ *everywhere.*
   - *(b)* $\mathrm{Supp}(g) \subseteq [0, 1]$.
   - *(c)* $\int_0^1 g(x)\, dx = 1$.
   - *(d)* g *is* m *times continuously differentiable and for all* $k \leq m$, $|g^{(k)}(x)| = O((2m)^{k+1})$.

   The function $\frac{1}{K} g_{(K)}(x)$ *satisfies* $\mathrm{Supp}(g_{(K)}) \subseteq [0, K]$, $\int_0^K g_{(K)}\, dx = 1$, *and for* $k \leq m$, $g_{(K)}^{(k)}(x) = O\left(\left(\frac{2m}{K}\right)^{k+1}\right)$.

2. *There exists a function $G : \mathbb{R} \to \mathbb{R}$ with the following properties.*
   - *(a)* G *is nondecreasing.*
   - *(b)* $G(x) = 0$ *for* $x \leq 0$.
   - *(c)* $G(x) = 1$ *for* $x \geq 1$.
   - *(d)* G *is* $m + 1$ *times continuously differentiable and for all* $k \leq m$, $G^{(k)}(x) = O((2m)^k)$.

3. *There exists a function $b : \mathbb{R} \to \mathbb{R}$ with the following properties:*
   - *(a)* $\mathrm{Supp}(b) \subseteq [-2, 2]$.
   - *(b)* $b(x) = 1$ *for* $x \in [-1, 1]$.
   - *(c)* b *is is* $m + 1$ *times continuously differentiable and for all* $k \leq m$, $b^{(k)}(x) = O((2m)^k)$.

   The function $b_{(K)}$ *satisfies* $\mathrm{Supp}(b_{(K)}) \subseteq [-2K, 2K]$, $b_{(K)}(x) = 1$ *for* $x \in [-K, K]$, *and* $b_{(K)}^{(m)}(x) = O\left(\left(\frac{2m}{K}\right)^k\right)$.

Proof Take

$$g(x) = \begin{cases} C_m 4^{m+1} x^{m+1}(1-x)^{m+1}, & x \in [0,1] \\ 0, & \text{else.} \end{cases}$$

where C_m is chosen so that $\int_0^1 g(x)\, dx = 1$. Note that $x(1-x) \leq \frac{1}{4}$ so $g(x) \leq C_m$ and

$$1 = \int_0^1 g(x)\, dx \leq C_m \tag{41}$$

$$1 = \int_0^1 g(x)\, dx \geq \int_{\frac{1}{2} - \frac{1}{2\sqrt{m}}}^{\frac{1}{2} + \frac{1}{2\sqrt{m}}} C_m 4^{m+1} x^{m+1}(1-x)^{m+1}\, dx \tag{42}$$

$$\geq \frac{1}{\sqrt{m}} C_m 4^{m+1} \left(\frac{1}{2} + \frac{1}{2\sqrt{m}}\right)^{m+1} \left(\frac{1}{2} - \frac{1}{2\sqrt{m}}\right)^{m+1} \tag{43}$$

$$\geq \frac{1}{\sqrt{m}} C_m \left(1 - \frac{1}{m}\right)^{m+1} \tag{44}$$

$$\geq \frac{C_m}{2e\sqrt{m}} \tag{45}$$

so $1 \le C_m \le 2e\sqrt{m}$.

Now, note that for functions u, v,

$$(uv)^{(k)} = \sum_{j=0}^{k} \binom{k}{j} u^{(j)} v^{(k-j)}. \tag{46}$$

Applying this to x^{m+1} and $(1-x)^{m+1}$ and gives that for $0 \le x \le 1$, $k \le m$,

$$\begin{aligned} |g^{(k)}(x)| &\le C_m \sqrt{m} \sum_{j=0}^{k} \binom{k}{j} (m+1)^j (m+1)^{k-j} && (47)\\ &\le O(m(2(m+1))^k) && (48)\\ &= O((2m)^{k+1}). && (49) \end{aligned}$$

For the second part, take $F(x) = \int_{-\infty}^{x} f(t)\,dt$. The normalization $\int_0^1 f(x)\,dx = 1$ ensures $F(x) = 1$ for $x \ge 1$, and for $k \le m$, $F^{(k+1)}(x) = f^{(k)}(x) = O((2m)^k)$.

For the third part, define

$$b(x) = \begin{cases} 0, & |x| > 2 \\ F(2-|x|), & 1 \le |x| \le 2 \\ 1, & |x| < 1. \end{cases}$$

For the rescaled functions, just note that for any function f, $f_{(K)}^{(k)}(x) = \frac{1}{K^k} f^{(k)}\left(\frac{x}{K}\right)$. ■

Appendix E. Omitted Proofs in Section 4

Theorem E.1 (Theorem 4.2 restated) *If f is differentiable, then for any g such that* $\mathrm{Supp}(g) \subseteq rB_n$ *and $g, \widehat{g} \in L^1(\mathbb{R}^n)$,*

$$C_{f,rB_n} \ge r \frac{\int_{\mathbb{R}^n} |\widehat{(\nabla f) g}(\omega)|\,d\omega}{\int_{\mathbb{R}^n} |\widehat{g}(\omega)|\,d\omega}$$

Proof Let $B = rB_n$. We have

$$\begin{aligned} C_{f,B} &= \inf_{F|_B = f} \int_{\mathbb{R}^n} \|\omega\|_B \, |\widehat{F}(\omega)|\,d\omega && (50)\\ &= r \inf_{F|_B = f} \int_{\mathbb{R}^n} \|\omega\|_2 \, |\widehat{F}(\omega)|\,d\omega && (51)\\ &= r \inf_{F|_B = f} \int_{\mathbb{R}^n} \left\|\widehat{\nabla F}(\omega)\right\|_2 d\omega. && (52) \end{aligned}$$

Young's inequality and Theorem A.4 give

$$\begin{aligned} \int_{\mathbb{R}^n} \left\|\widehat{\nabla F}(\omega)\right\|_2 d\omega \int_{\mathbb{R}^n} |\widehat{g}(\omega)|\,d\omega &\ge \int_{\mathbb{R}^n} \left\|(\widehat{\nabla F} * \widehat{g})(\omega)\right\|_2 d\omega && (53)\\ &= \int_{\mathbb{R}^n} \left\|\widehat{(\nabla F) g}(\omega)\right\|_2 d\omega && (54)\\ &= \int_{\mathbb{R}^n} \left\|\widehat{(\nabla f) g}(\omega)\right\|_2 d\omega. && (55) \end{aligned}$$

where the last step uses the fact that $\mathrm{Supp}(g) \subseteq rB_n$, so $(\nabla F)g = (\nabla f)g$. Then

$$\int_{\mathbb{R}^n} \left\|\widehat{\nabla F}(\omega)\right\|_2 \, d\omega \geq \frac{\int_{\mathbb{R}^n} \left\|\widehat{(\nabla f)g}(\omega)\right\|_2 \, d\omega}{\int_{\mathbb{R}^n} |\widehat{g}(\omega)| \, d\omega}. \tag{56}$$

■

E.1. f is not Barron

In this section we prove Lemma 4.3. We first prove the function g we choose gives a small denominator in the lowerbound equation.

Lemma E.2 *For $n \equiv 3 \pmod 4$,*

$$\int_{\mathbb{R}^n} \|\widehat{g}(\omega)\| \, d\omega \leq O((5eC_2)^{\frac{n}{2}}).$$

To prove this we will need bound certain combinations of derivatives of a radial function.

Lemma E.3 *Let $f\colon \mathbb{R}^n \to \mathbb{R}^n$ be a radial function with $f(x) = f_1(\|x\|)$. Then for $k \in \mathbb{N}$, $1 \leq k \leq \frac{n}{4} + 1$,*

$$((I-\Delta)^k f)(x) = \sum_{\substack{0 \leq i \leq 2k, 0 \leq j \leq \max\{0, 2k-1\} \\ i+j \leq 2k}} \frac{c_{i,j} n^j f_1^{(i)}(r)}{r^j}, \quad r = \|x\| \tag{57}$$

for some $c_{i,j}$ with $\sum_{i,j} |c_{i,j}| \leq 5^k$.

Here, $(I-\Delta)f$ denotes $f - \Delta f$.

Proof We proceed by induction. The case $k = 0$ is just $f(x) = f_1(r)$. Suppose the statement is true for a given $k \leq \frac{n}{4}$; we show it for $k+1$. Let $(I-\Delta)^k f$ be given by (57). We use the formula for the Laplacian of a radial function,

$$\Delta f(x) = \frac{n-1}{r} f_1'(r) + f_1''(r). \tag{58}$$

For ease of notation, in the below the arguments of f and f_1, which are x and r, are omitted. Then using (58) and the product rule,

$$(I-\Delta)^{k+1} f = \sum_{\substack{0 \leq i \leq 2k, 0 \leq j \leq \max\{0, 2k-1\} \\ i+j \leq 2k}} c_{i,j} n^j \left(\frac{1}{r^j} f_1^{(i)} + \frac{n-1}{r} \left(\frac{j}{r^{j+1}} f_1^{(i)} - \frac{1}{r^j} f_1^{(i+1)} \right) \right. \tag{59}$$

$$\left. + \left(-\frac{j(j+1)}{r^{j+2}} f_1^{(i)} + \frac{2j}{r^{j+1}} f_1^{(i+1)} - \frac{1}{r^j} f_1^{(i+2)} \right) \right) \tag{60}$$

The largest derivative of f_1 increases by 2 and the power of r increases by 2, except when $k = 0$, when the power increases by 1 (from (58)). Write this as

$$\sum_{\substack{0 \le i \le 2(k+1), 0 \le j \le 2k+1 \\ i+j \le 2(k+1)}} \frac{c'_{i,j} n^j f_1^{(i)}}{r^j}.$$

A term is identified by the order $f^{(i)}$ that appears and the power $\frac{1}{r^j}$ that appears. For example, the term $c_{i,j} n^j \frac{n-1}{r} \frac{j}{r^{j+1}} f_1^{(i)} = c_{i,j} n^{j+2} \frac{(n-1)j}{n^2} \frac{1}{r^{j+2}} f_1^{(i)}$ in (59) will contribute $c_{i,j} \frac{(n-1)j}{n^2}$ to $c'_{i,j+2}$. Noting $k \le \frac{n}{4}$ implies $2k \le \frac{n}{2}$, we have

$$\sum_{i,j} |c'_{i,j}| \le \sum_{\substack{0 \le i \le 2k, 0 \le j \le \max\{0, 2k-1\} \\ i+j \le 2k}} |c_{i,j}| \left(1 + \frac{(n-1)j}{n^2} + \frac{n-1}{n} + \frac{j(j+1)}{n^2} + \frac{2j}{n} + 1\right) \tag{61}$$

$$\le \sum_{i,j} |c_{i,j}| \left(1 + \frac{1}{2} + 1 + \frac{1}{4} + 1 + 1\right) \tag{62}$$

$$\le 5 \sum_{i,j} |c_{i,j}|. \tag{63}$$

This completes the induction step and proves the theorem. ■

Proof [Proof of Lemma E.2] By Lemma A.6 with $k = \frac{n+1}{2}$,

$$\int_{\mathbb{R}^n} \|\widehat{g}(\omega)\| \, d\omega \le \left(\frac{\Gamma\left(\frac{1}{2}\right)}{2^n \pi^{\frac{n}{2}} \Gamma\left(\frac{n+1}{2}\right)} \right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^n} [(I - \Delta)^{\frac{n+1}{4}} g(x)]^2 \, dx \right)^{\frac{1}{2}}. \tag{64}$$

Note $[(I-\Delta)^{\frac{n+1}{4}} g(x)]^2$ is nonzero only on $2K_2 B_n$. Then letting $c_{i,j}$ be as in Lemma E.3 with $k = \frac{n+1}{4}$, we have

$$(I - \Delta)^{\frac{n+1}{4}} g(x) = \sum_{\substack{0 \le i \le \frac{n+1}{2}, 0 \le j \le \frac{n-1}{2} \\ i+j \le \frac{n+1}{2}}} \frac{c_{i,j} n^j g_1^{(i)}(r)}{r^j}, \quad r = \|x\| \tag{65}$$

We separate out the one term $g_1(r)$, and bound the derivatives noting that g_1 was defined using the bump function $b_{(K_2)}$ in Lemma D.1. Note that $g_1^{(i)} = 0$ for $r < K_2$, so we can take $r \ge K_2$ in the sum.

$$\begin{aligned}
|(I-\Delta)^{\frac{n+1}{4}} g(x)| &\le g_1(r) + \sum_{\substack{1\le i\le \frac{n+1}{2}, 0\le j\le \frac{n-1}{2}\\ i+j\le \frac{n+1}{2}}} |c_{i,j}| \frac{n^j |g_1^{(i)}(r)|}{r^j} \qquad (66)\\
&\le g_1(r) + \sum_{\substack{1\le i\le \frac{n+1}{2}, 0\le j\le \frac{n-1}{2}\\ i+j\le \frac{n+1}{2}}} |c_{i,j}| \frac{n^j O\left(\frac{(n+1)^i}{(C_2 n)^i}\right)}{(C_2 n)^j} \qquad (67)\\
&= O(4^{\frac{n+1}{4}}). \qquad (68)\\
& \qquad (69)
\end{aligned}$$

Noting that the volume of $2K_2B_n$ is $\frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)}(2K_2)^n$,

$$\begin{aligned}
\left(\int_{\mathbb{R}^n} [(I-\Delta)^{\frac{n+1}{4}} g(x)]^2\,dx\right)^{\frac{1}{2}} &= O\left(\left(\frac{\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}+1\right)}(2K_2)^n \left(5^{\frac{n+1}{4}}\right)^2\right)^{\frac{1}{2}}\right) \qquad (70)\\
&= O\left(\left(\frac{\pi^{\frac{n}{2}} 2^n C_2^n n^n}{\Gamma(\frac{n}{2}+1)}\right)^{\frac{1}{2}} 5^{\frac{n+1}{4}}\right). \qquad (71)
\end{aligned}$$

Combining (64) and (71) and using Stirling's approximation $\Gamma(n+1)\sim\sqrt{2\pi n}\left(\frac{n}{e}\right)^n$ gives

$$\begin{aligned}
\int_{\mathbb{R}^n} \|\widehat{g}(\omega)\|\,d\omega &\le O\left(\frac{C_2^{\frac{n}{2}} n^{\frac{n}{2}} 5^{\frac{n+1}{4}}}{\Gamma\left(\frac{n+1}{2}\right)^{\frac{1}{2}} \Gamma\left(\frac{n}{2}+1\right)^{\frac{1}{2}}}\right) \qquad (72)\\
&= O\left((5eC_2)^{\frac{n}{2}}\right). \qquad (73)
\end{aligned}$$

■

Now we are ready to bound the numerator and finish the proof.

Lemma E.4 *For f defined as in Section 4.1, $n\equiv 3 \pmod 4$, and constants C_1, C_2, C_3 such that $C_1C_3 \ge \frac{3}{2}$, $C_2 > C_1 \ge 1$, $C_3 \ge 1$,*

$$C_{f,2K_3B_n} = \Omega\left(2^{-n} C_1^{\frac{n}{2}-3} C_3^{\frac{n}{2}} C_2^{-\left(\frac{n}{2}-1\right)} n^{\frac{1}{2}}\right).$$

In particular, this is exponentially large if we choose C_3 large enough (i.e. if we make f vary sharply enough).

Proof For $\|\omega\| = K_3$, by (17), (18), and Lemma B.1,

$$\begin{aligned}
\widehat{f}(\omega) &= \frac{1}{2\pi}\left(\frac{1}{2\pi K_3}\right)^{\frac{n}{2}-1} \int_{K_1}^{K_1+\varepsilon} r^{\frac{n}{2}-1} f_1(r) J_{\frac{n}{2}-1}(K_3 r)\,dr \qquad (74)\\
&\ge \frac{1}{2\pi}\left(\frac{1}{2\pi K_3}\right)^{\frac{n}{2}-1} \int_{K_1}^{K_1+\varepsilon} r^{\frac{n}{2}-1} f_1(r) \left(\sqrt{\frac{2}{\pi K_3 r}}\frac{1}{\sqrt{2}} - (K_3 r)^{-\frac{3}{2}}\right) dr \qquad (75)\\
&\ge \frac{1}{2\pi}\left(\frac{K_1}{2\pi K_3}\right)^{\frac{n-3}{2}} \sqrt{\frac{1}{\pi}}(1-o(1)) \qquad (76)
\end{aligned}$$

where in the last step we used $\int_{K_1}^{K_1+\varepsilon} f_1(r) = 1$. Now we show that $\widehat{f}$ is also large for $\|\omega\| \approx K_3$. Let ω, ω_0 be such that $\|\omega_0\| = K_3$ and $\omega \geq \omega_0$. Then using the fact that $J_{\frac{n}{2}-1}$ is 1-Lipschitz for $x \geq 3\left(\frac{n}{2}-1\right)$ (Lemma B.2) and $K_3 K_1 \geq C_3 C_1 n \geq \frac{3n}{2}$,

$$\begin{aligned}
|\widehat{f}(\omega) - \widehat{f}(\omega_0)| &\leq \frac{1}{2\pi}\left(\frac{1}{2\pi K_3}\right)^{\frac{n}{2}-1} \int_{K_1}^{K_1+\varepsilon} r^{\frac{n}{2}-1} f_1(r) |J_{\frac{n}{2}-1}(\|\omega\|\, r) - J_{\frac{n}{2}}(K_3 r)|\, dr && (77)\\
&\leq \frac{1}{2\pi}\left(\frac{1}{2\pi K_3}\right)^{\frac{n}{2}-1} \int_{K_1}^{K_1+\varepsilon} r^{\frac{n}{2}-1} f_1(r) r (\|\omega\| - K_3)\, dr && (78)\\
&\leq \frac{1}{2\pi}\left(\frac{1}{2\pi K_3}\right)^{\frac{n}{2}-1} (K_1+\varepsilon)^{\frac{n}{2}} (\|\omega\| - K_3) && (79)\\
&= O\left(\left(\frac{K_1}{2\pi K_3}\right)^{\frac{n}{2}-1} K_1^{\frac{3}{2}} K_3^{\frac{1}{2}} (\|\omega\| - K_3)\right) && (80)
\end{aligned}$$

By (76) and (80), for $n \geq 3$, there exists δ such that for all $\|\omega\| \in \left[K_3, K_3 + \frac{\delta}{K_1^{3/2} K_3^{1/2}}\right]$,

$$|\widehat{f}(\omega)| = \Omega\left(\left(\frac{K_1}{2\pi K_3}\right)^{\frac{n-3}{2}}\right) \tag{81}$$

Then using the fact that the surface area of a sphere in $\mathbb{R}^n$ is $\frac{2\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)}$,

$$\begin{aligned}
\int_{\mathbb{R}^n} \|\omega\|\, |\hat{f}(\omega)|\, d\omega &= \int_{K_3 \leq \|\omega\| \leq K_3 + \frac{\delta}{K_1^{3/2}}} \Omega\left(\left(\frac{K_1}{2\pi K_3}\right)^{\frac{n-3}{2}}\right) d\omega && (82)\\
&= \Omega\left(\frac{\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} K_3^{n-1} \frac{\delta}{K_1^{3/2} K_3^{1/2}} \left(\frac{K_1}{2\pi K_3}\right)^{\frac{n-3}{2}}\right) && (83)\\
&= \Omega\left(\frac{1}{\Gamma\left(\frac{n}{2}\right)} K_3^{\frac{n}{2}} K_1^{\frac{n}{2}-3} 2^{-\frac{n}{2}}\right) && (84)\\
&= \Omega\left(\left(\frac{2e}{n-2}\right)^{\frac{n}{2}-1} (C_3 n^{\frac{1}{2}})^{\frac{n}{2}} (C_1 n^{\frac{1}{2}})^{\frac{n}{2}-3} 2^{-\frac{n}{2}}\right) && (85)\\
&= \Omega(C_1^{\frac{n}{2}-3} C_3^{\frac{n}{2}} n^{-\frac{1}{2}} e^{\frac{n}{2}}). && (86)
\end{aligned}$$

Note $K_2 = C_2 n > C_1\sqrt{n} + \varepsilon = K_1 + \varepsilon$. Then $g = 1$ on the support of f, so $(\nabla f) g = \nabla f$ and

$$\begin{aligned}
\int_{\mathbb{R}^n} \left\|\widehat{(\nabla f) g}(\omega)\right\| d\omega &= \int_{\mathbb{R}^n} \left\|\widehat{\nabla f}(\omega)\right\| d\omega && (87)\\
&= \Omega(C_1^{\frac{n}{2}-3} C_3^{\frac{n}{2}} n^{-\frac{1}{2}} e^{\frac{n}{2}}). && (88)
\end{aligned}$$

Then by Lemma E.2,

$$\begin{aligned}
C_{f, 2K_2 B_n} &\geq 2K_2 \frac{\int_{\mathbb{R}^n} \left\|\widehat{(\nabla f) g}(\omega)\right\| d\omega}{\int_{\mathbb{R}^n} |\widehat{g}(\omega)|\, d\omega} && (89)\\
&= 2K_2 \frac{\Omega(C_1^{\frac{n}{2}-3} C_3^{\frac{n}{2}} n^{-\frac{1}{2}} e^{\frac{n}{2}})}{O((5eC_2)^{\frac{n}{2}})} = \Omega\left(5^{-\frac{n}{2}} C_1^{\frac{n}{2}-3} C_3^{\frac{n}{2}} C_2^{-\left(\frac{n}{2}-1\right)} n^{\frac{1}{2}}\right). && (90)
\end{aligned}$$

■

E.2. h is a composition of Barron functions

In this section we proof Lemma 4.4. In order to do that, let us first define the following set of functions:

Definition E.5 *Define*

$$\Gamma(A, C) := \left\{ f \colon \mathbb{R}^n \to \mathbb{R} : \int_{\mathbb{R}^n} |\widehat{f}(\omega)| \, d\omega \le A, \int_{\mathbb{R}^n} \|\omega\| \, |\widehat{f}(\omega)| \, d\omega \le C \right\}$$

Barron functions have many nice properties:

Proposition E.6 (Properties of Barron constant)

1. *(Subadditivity, (Barron, 1993, §IV.3)) For any set B,*

$$C_{\sum_i \beta_i f_i, B} \le \sum_i |\beta_i| C_{f_i, B}.$$

2. *(Ridge functions, (Barron, 1993, §IV.7)) Suppose $f = h(\langle a, x\rangle)$, where $h : \mathbb{R} \to \mathbb{R}$ is a 1-dimensional function and $\|a\|_2 = 1$. Then*

$$C_{f, rB_n} \le C_{h, [-r, r]}.$$

3. *(Powers, (Barron, 1993, §IV.12)) If $g : \mathbb{R} \to \mathbb{R}$, $g \in \Gamma(a, c)$, then $g(x)^k \in \Gamma(a^k, ka^{k-1}c)$.*

4. *The function $f(x) = x$ has an extension h agreeing with x on $[-r, r]$, which satisfies $h(x) \in \Gamma(O(r^{\frac{3}{2}}), O(r^{\frac{1}{2}}))$.*

Proof We show (4). Choose a bump function b as in Lemma D.1 for $m = 2$. Consider the extension $h(x) = xb_{(r)}(x) = xb\left(\frac{x}{r}\right)$ which is supported on $[-2r, 2r]$. Because b, b', b'' are all bounded by a constant, on $[-2r, 2r]$,

$$|h(x)| \le x \tag{91}$$

$$|h'(x)| = |b_{(r)}(x) + xb'_{(r)}(x)| \le 1 + O\left(\frac{x}{r}\right) \tag{92}$$

$$|h''(x)| = |2b'_{(r)}(x) + b''_{(r)}(x)| \le O\left(\frac{x}{r}\right) + O\left(\frac{1}{r^2}\right). \tag{93}$$

Then by Lemma A.6(2),

$$\int_{-\infty}^{\infty} |\widehat{h}(\omega)| \, d\omega \le 2^{-\frac{1}{2}} \left(\int_{-r}^{r} |h(x)|^2 + |h'(x)|^2 \, dx \right)^{\frac{1}{2}} \le O(r^{\frac{3}{2}}) \tag{94}$$

$$\int_{-\infty}^{\infty} |\omega \widehat{h}(\omega)| \, d\omega \le 2^{-\frac{1}{2}} \left(\int_{-r}^{r} |h'(x)|^2 + |h''(x)|^2 \, dx \right)^{\frac{1}{2}} \le O(r^{\frac{1}{2}}). \tag{95}$$

■

Proof [Proof of Theorem 4.4] By Proposition E.6(4) and (3), the 1-dimensional function $y \mapsto y^2$ has an extension $k(y)$ with $k(y) \in \Gamma(O(r^3), O(r^2))$. Thus, $C_{y^2,[-r,r]} \le r \int_{-\infty}^{\infty} \|\omega\| \, |\widehat{k}(\omega)| \, d\omega = O(r^3)$.

Because $x_i^2 : \mathbb{R}^n \to \mathbb{R}$ is the composition of the projection $x \mapsto \langle e_i, x \rangle$ and the 1-dimensional function $y \mapsto y^2$ and , by (2),

$$C_{x_i^2, rB_n} \le C_{y^2,[-r,r]} \le O(r^3)$$

By (1), because $\|x\|^2 = \sum_{i=1}^n x_i^2$,

$$C_{\|x\|^2, rB_n} \le O(nr^3).$$

Now consider the function $h(y) := f_1(\sqrt{y})$. We have, noting this is nonzero only for $x \in [K_1^2, (K_1+\varepsilon)^2]$, and $f_1^{(i)}(\sqrt{y}) = O(K_3^{i+1})$,

$$h'(y) = \frac{1}{2y^{\frac{1}{2}}} f_1(\sqrt{y}) + f_1'(\sqrt{y}) = O\left(\left(\frac{K_3}{K_1}\right) + K_3^2\right) \tag{96}$$

$$h''(y) = \frac{1}{4y^{\frac{3}{2}}} f_1(\sqrt{y}) + \frac{1}{4y} f_1'(\sqrt{y}) + \frac{1}{2y^{\frac{1}{2}}} f_1''(\sqrt{y}) = O\left(\frac{K_3}{K_1^3} + \frac{K_3^2}{K_1^2} + \frac{K_3^3}{K_1}\right). \tag{97}$$

Using $C_3 < C_1$ we have $|h'|^2 + |h''|^2 = O(K_3^4)$. Thus by Lemma A.6,

$$\int_0^\infty |\omega \widehat{h}(\omega)| \, d\omega = \left(\int_{K_1^2}^{(K_1+\varepsilon)^2} O\left(K_3^4\right)\right)^{\frac{1}{2}} = O\left(\left(\frac{K_1}{K_3} O(K_3^4)\right)^{\frac{1}{2}}\right) = O\left(K_1^{\frac{1}{2}} K_3^{\frac{3}{2}}\right).$$

Thus $f_1(\sqrt{x})$ is $O(sC_1^{\frac{1}{2}} C_3^{\frac{3}{2}} n^2)$-Barron on $[-s, s]$. ■