A Function Space View of Bounded Norm Infinite Width ReLU Nets: The Multivariate Case

Anonymous authors
Paper under double-blind review

Abstract

A key element of understanding the efficacy of overparameterized neural networks is characterizing how they represent functions as the number of weights in the network approaches infinity. In this paper, we characterize the norm required to realize a function $f : \mathbb{R}^d \to \mathbb{R}$ as a single hidden-layer ReLU network with an unbounded number of units (infinite width), but where the Euclidean norm of the weights is bounded, including precisely characterizing which functions can be realized with finite norm. This was settled for univariate functions $f : \mathbb{R} \to \mathbb{R}$ in Savarese et al. (2019), where it was shown that the required norm is determined by the L^1-norm of the second derivative of the function. We extend the characterization to multi-variate functions ($d \geq 2$, *i.e.*, multiple input units), relating the required norm to the L^1-norm of the Radon transform of a $(d+1)/2$-power Laplacian of the function. This characterization allows us to show that all functions in Sobolev spaces $W^{s,1}(\mathbb{R}^d)$, $s \geq d+1$, can be represented with bounded norm, to calculate the required norm for several specific functions, and to obtain a depth separation result. These results have important implications for understanding generalization performance and the distinction between neural networks and more traditional kernel learning.

1 Introduction

It has been argued for a while, and is becoming increasingly apparent in recent years, that in terms of complexity control and generalization in neural network training, “the size [magnitude] of the weights is more important then the size [number of weights or parameters] of the network” (Bartlett, 1997; Neyshabur et al., 2014; Zhang et al., 2016). That is, inductive bias and generalization are not achieved by limiting the size of the network, but rather by explicitly or implicitly (Neyshabur et al., 2014; 2017; Gunasekar et al., 2018) controlling the magnitude of the weights.

In fact, since networks used in practice are often so large that they can fit any function (any labels) over the training data, it is reasonable to think of the network as virtually infinite-sized, and thus able to represent essentially all functions. Training and generalization ability then rests on fitting the training data while controlling, either explicitly or implicitly, the magnitude of the weights. That is, training searches over all functions, but seeks functions with small *representational cost*, given by the minimal weight norm required to represent the function. This “representational cost of a function” is the actual inductive bias of learning—the quantity that defines our true model class, and the functional we are actually minimizing in order to learn. Understanding learning with overparameterized (virtually infinite) networks thus rests on understanding this “representational cost”, which is the subject of our paper.

We can also think of understanding the representational cost as asking an approximation theory question: what functions can we represent, or approximate, with our de facto model class, namely the class of functions representable with small magnitude weights? There has been much celebrated work studying approximation in terms of the network *size*, *i.e.*, asking how many units are necessary in order to approximate a target function (Hornik et al., 1989; Cybenko, 1989; Barron, 1993; Pinkus, 1999). But if complexity is actually controlled by the norm of the weights, and thus our true model class is defined by the magnitude of the weights, we should instead ask how large a norm is necessary in order to capture a target function. This revised view of approximation theory should also change how we view issues such as depth separation: rather then asking how increasing depth can reduce the *number* of units required to fit a function, we should instead ask how increasing depth can reduce the *norm* required, *i.e.*, how the representational cost we study changes with depth.

Our discussion above directly follows that of Savarese et al. (2019), who initiated the study of the representational cost in term of weight magnitude. Savarese et al. considered two-layer (*i.e.*, single hidden layer) ReLU networks, with an unbounded (essentially infinite) number of units, and where the overall Euclidean norm (sum of squares of all the weights) is controlled. Infinite width networks of this sort have been studied from various perspectives by *e.g.*, Bengio et al. (2006); Neyshabur et al. (2015); Bach (2017); Mei et al. (2018). For univariate functions $f : \mathbb{R} \to \mathbb{R}$, corresponding to networks with a single one-dimensional input and a single output, Savarese et al. obtained a crisp and precise characterization of the representational cost, showing that minimizing overall Euclidean norm of the weights is equivalent to fitting a function by controlling:

$$\max\left(\int |f''(x)|dx, |f'(-\infty) + f'(+\infty)|\right). \tag{1}$$

While this is an important first step, we are of course interested also in more than a single one-dimensional input. In this paper we derive the representational cost for any function $f : \mathbb{R}^d \to \mathbb{R}$ in any dimension d. Roughly speaking, the cost is captured by:

$$\|f\|_{\mathcal{R}} \dot{\approx} \|\mathcal{R}\{\Delta^{(d+1)/2} f\}\|_1 \approx \|\partial_b^{d+1} \mathcal{R}\{f\}\|_1 \tag{2}$$

where $\mathcal{R}$ is the Radon transform, Δ is the Laplacian, and ∂_b is a partial derivative w.r.t. the offset in the Radon transform (see Section 3 for an explanation of the Radon transform). This characterization is rigorous for odd dimensions d and for functions where the above expressions are classically well-defined (*i.e.*, smooth enough such that all derivatives are finite, and the integrand in the Radon transform is integrable). But for many functions of interest these quantities are not well-defined classically. Instead, in Definition 1, we use duality to rigorously define a semi-norm $\|f\|_{\mathcal{R}}$ that captures the essence of the above quantities and is well-defined (though possibly infinite) for any f in any dimension. We show that $\|f\|_{\mathcal{R}}$ precisely captures the representational cost of f, and in particular is finite if and only if f can be approximated arbitrarily well by a bounded norm, but possibly unbounded width, ReLU network. Our precise characterization applies to an architecture with unregularized bias terms (as in Savarese et al. (2019)) and a single unregularized linear unit—otherwise a correction accounting for a linear component is necessary, similar but more complex than the term $|f'(-\infty) + f'(+\infty)|$ in the univariate case, *i.e.*, (1).

As we uncover, the characterization of the representational cost for multivariate functions is unfortunately not as simple as the characterization (1) in the univariate case, where the Radon transform degenerates. Nevertheless, it is often easy to evaluate, and is a powerful tool for studying the representational power of bounded norm ReLU networks. Furthermore, as detailed in Section 5.5, we are unaware of any kernel function for which the associated RKHS norm is the same as (2); *i.e.*, training neural networks is fundamentally different from kernel learning. In particular, using our characterization we show the following:

- All sufficiently smooth functions have finite representational cost, but the necessary degree of smoothness depends on the dimension. In particular, all functions in the Sobolev space $W^{d+1,1}(\mathbb{R}^d)$, *i.e.*, when all derivatives up to order $d+1$ are L^1-bounded, have finite representational cost, and this cost can be bounded using the Sobolev norm. (Section 5.1)
- We calculate the representational cost of radial "bumps", and show there are bumps with finite support that have finite representational cost in all dimensions. The representational cost increases as $1/\varepsilon$ for "sharp" bumps of radius ε (and fixed height). (Section 5.2)
- In dimensions greater than one, piecewise linear functions with bounded support have infinite representational cost (*i.e.*, cannot be represented with a bounded norm, even with infinite networks). (Section 5.3)
- We obtain a depth separation in terms of norm: we demonstrate a function in two dimensions that is representable using a depth three ReLU network (*i.e.*, with two hidden layers) with small finite norm, but cannot be represented by any bounded-norm depth two (single hidden layer) ReLU network. As far as we are aware, this is the first depth separation result in terms of the norm required for representation. (Section 5.4)

1.1 Related Work

Although the focus of most previous work on approximation theory for neural networks was on the number of units, the norm of the weights was often used as an intermediate step. However, this use does not provide an exact characterization of the representational cost, only a (often very loose) upper bound, and in particular does not allow for depth separation results where a *lower bound* is needed. See Savarese et al. (2019) for a detailed discussion, *e.g.*, contrasting with the work of Barron (1993; 1994).

The connection between the Radon transform and two-layer neural networks was previously made by Carroll & Dickinson (1989) and Ito (1991), who used it to obtain constructive approximations when studying approximation theory in terms of network size (number of units) for threshold and sigmoidal networks. This connection also forms the foundation of ridgelet transform analysis of functions Candès & Donoho (1999); Candès (1999). More recently, Sonoda & Murata (2017) used ridgelet transform analysis to study the approximation properties of two-layer neural networks with unbounded activation functions, including the ReLU.

While working on this manuscript, we learned through discussions with Matus Telgarsky of his related parallel work. In particular, Telgarsky obtained a calculation formula for the norm required to represent a radial function, paralleling our calculations in Section 5.2, and used it to show that sufficiently smooth radial functions have finite norm in any dimension, and studied how this norm changes with dimension.

2 Infinite Width ReLU Networks

We repeat here the discussion of Savarese et al. (2019) defining the representational cost of infinite-width ReLU networks, with some corrections and changes that we highlight.

Consider the collection of all two-layer networks having an unbounded number of rectified linear units (ReLUs), *i.e.*, all $g_\theta : \mathbb{R}^d \to \mathbb{R}$ defined by

$$g_\theta(\boldsymbol{x}) = \sum_{i=1}^{k} a_i[\boldsymbol{w}_i^\top \boldsymbol{x} - b_i]_+ + c, \quad \text{for all } \ \boldsymbol{x} \in \mathbb{R}^d \tag{3}$$

with parameters $\theta = (k, \boldsymbol{W} = [\boldsymbol{w}_1, ..., \boldsymbol{w}_k], \boldsymbol{b} = [b_1, ..., b_k]^\top, \boldsymbol{a} = [a_1, ..., a_k]^\top, c)$, where the width $k \in \mathbb{N}$ is unbounded. Let Θ be the collection of all such parameter vectors θ. For any $\theta \in \Theta$ we let $C(\theta)$ be the sum of the squared Euclidean norm of the weights in the network excluding the bias terms, *i.e.*,

$$C(\theta) = \frac{1}{2}\left(\|\boldsymbol{W}\|_F^2 + \|\boldsymbol{a}\|^2\right) = \frac{1}{2}\sum_{i=1}^{k}\left(\|\boldsymbol{w}_i\|_2^2 + |a_i|^2\right), \tag{4}$$

and consider the minimal representation cost necessary to exactly represent a function $f \in \mathbb{R}^d \to \mathbb{R}$

$$R(f) := \inf_{\theta \in \Theta} C(\theta) \ \ s.t. \ \ f = h_\theta. \tag{5}$$

By the 1-homogeneity of the ReLU, it is shown in Neyshabur et al. (2014) (see also Appendix A of Savarese et al. (2019)) that minimizing $C(\theta)$ is the same as constraining the inner layer weight vectors $\{\boldsymbol{w}_i\}_{i=1}^k$ to be unit norm while minimizing the ℓ^1-norm of the outer layer weights $\boldsymbol{a}$. Therefore, letting Θ' be the collection of all $\theta \in \Theta$ with each $\boldsymbol{w}_i$ constrained to the unit sphere $\mathbb{S}^{d-1} := \{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\| = 1\}$, we have

$$R(f) = \inf_{\theta \in \Theta'} \|\boldsymbol{a}\|_1 \ \ s.t. \ \ f = h_\theta. \tag{6}$$

However, we see $R(f)$ is finite only if f is exactly realizable as a finite-width two layer ReLU network, *i.e.*, f must be a continuous piecewise linear function with finitely many pieces. Yet, we know that any continuous function can be approximated uniformly on compact sets by allowing the number of ReLU units to grow to infinity. Since we are not concerned with the number of units, only their norm, we modify our definition of representation cost to capture this larger space of functions, and define[1]

$$\overline{R}(f) := \lim_{\varepsilon \to 0} \left(\inf_{\theta \in \Theta'} C(\theta) \;\; s.t. \;\; |g_\theta(\boldsymbol{x}) - f(\boldsymbol{x})| \leq \varepsilon \;\; \forall \, \|\boldsymbol{x}\| \leq 1/\varepsilon \text{ and } g_\theta(\mathbf{0}) = f(\mathbf{0}) \right) \tag{7}$$

In words, $\overline{R}(f)$ is the minimal limiting representational cost among all sequences of networks converging to f uniformly (while agreeing with f at zero).

Intuitively, if $\overline{R}(f)$ is finite this means f is expressible as an "infinite-width" ReLU network whose outer-most weights are described by a density $\alpha(\boldsymbol{w}, b)$ defined over all weight and bias pairs $(\boldsymbol{w}, b) \in \mathbb{S}^{d-1} \times \mathbb{R}$. To make this intuition precise, let $M(\mathbb{S}^{d-1} \times \mathbb{R})$ denote the space of signed measures α defined on $(\boldsymbol{w}, b) \in \mathbb{S}^{d-1} \times \mathbb{R}$ with finite total variation norm $\|\alpha\|_1 = \int_{\mathbb{S}^{d-1}\times\mathbb{R}} d|\alpha|$ (*i.e.*, the analog of the L^1-norm for measures), and let $c \in \mathbb{R}$. Then we define the infinite-width two-layer ReLU network $h_{\alpha,c}$ (or "infinite-width net" for short) by[2]

$$h_{\alpha,c}(\boldsymbol{x}) := \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \left([\boldsymbol{w}^\top \boldsymbol{x} - b]_+ - [-b]_+ \right) d\alpha(\boldsymbol{w}, b) + c \tag{8}$$

We prove in Appendix D that $\overline{R}(f)$ is equivalent to

$$\overline{R}(f) = \min_{\alpha \in M(\mathbb{S}^{d-1}\times\mathbb{R}), c \in \mathbb{R}} \|\alpha\|_1 \;\; s.t. \;\; f = h_{\alpha,c} \tag{9}$$

Hence, learning an unbounded width ReLU network g_θ by fitting some loss functional $L(\cdot)$ while controlling the Euclidean norm of the weights $C(\theta)$ by minimizing

$$\min_{\theta \in \Theta} L(g_\theta) + \lambda C(\theta) \tag{10}$$

is effectively the same as learning a function f by controlling $\overline{R}(f)$:

$$\min_{f:\mathbb{R}^d \to \mathbb{R}} L(f) + \lambda \overline{R}(f), \tag{11}$$

In other words, $\overline{R}(f)$ captures the true inductive bias of learning with unbounded width ReLU networks with regularized weights. Our goal is then to calculate $\overline{R}(f)$ for any function $f : \mathbb{R}^d \to \mathbb{R}$, and in particular characterize when it is finite in order to understand what functions can be approximated arbitrarily well with bounded norm but unbounded width ReLU networks.

2.1 Simplification using an unregularized linear unit

Every two-layer ReLU network uniquely decomposes into the sum of a network with absolute value units plus a linear part[3]. As demonstrated by Savarese et al. (2019) in the 1D setting, the weights on the absolute value units typically determine the representational cost, with a correction term needed if the linear part has large weight. To allow for a cleaner formulation of the representation cost without this correction term, we consider adding in *one additional unregularized linear unit* $\boldsymbol{v}^\top \boldsymbol{x}$ (similar to a "skip connection") to "absorb" any representational cost due to the linear part.

Namely, for any $\theta \in \Theta$ and $\boldsymbol{v} \in \mathbb{R}^d$ we define the class of unbounded with two-layer ReLU networks $g_{\theta,\boldsymbol{v}}$ with a linear unit by $g_{\theta,\boldsymbol{v}}(\boldsymbol{x}) = g_\theta(\boldsymbol{x}) + \boldsymbol{v}^\top \boldsymbol{x}$ with g_θ as defined in (3), and associate $g_{\theta,\boldsymbol{v}}$ with the same weight norm $C(\theta)$ as defined in (4) (*i.e.*, we exclude the norm of the weight $\boldsymbol{v}$ from the cost). We then define a new representational cost $\overline{R}_1(f)$ for this class of networks by

$$\overline{R}_1(f) := \lim_{\varepsilon \to 0} \left(\inf_{\theta \in \Theta', \boldsymbol{v} \in \mathbb{R}^d} C(\theta) \;\; s.t. \;\; |g_{\theta,\boldsymbol{v}}(\boldsymbol{x}) - f(\boldsymbol{x})| \leq \varepsilon \;\; \forall \, \|\boldsymbol{x}\| \leq 1/\varepsilon \text{ and } g_\theta(\mathbf{0}) = f(\mathbf{0}) \right). \tag{12}$$

[1] Our definition of $\overline{R}(f)$ differs from the one given in Savarese et al. (2019). We require $|g_\theta(\boldsymbol{x}) - f(\boldsymbol{x})| \leq \varepsilon$ on the ball of radius $1/\varepsilon$ rather than all of $\mathbb{R}^d$, and we additionally require $g_\theta(\mathbf{0}) = f(\mathbf{0})$. These modifications are needed to ensure (7) and (9) are equivalent. Also, we note the choice of zero in the condition $g_\theta(\mathbf{0}) = f(\mathbf{0})$ is arbitrary and can be replaced with any point $\boldsymbol{x}_0 \in \mathbb{R}^d$.

[2] Our definition of $h_{\alpha,c}$ also differs from the one given in Savarese et al. (2019). To ensure the integral is well-defined, we include the additional $-[-b]_+$ term in the integrand. See Remark 1 in the Appendix for more discussion on this point.

[3] Such a decomposition follows immediately from the identity $[t]_+ = \frac{1}{2}(|t| + t)$

Likewise, for all $\alpha \in M(\mathbb{S}^{d-1} \times \mathbb{R})$, $\boldsymbol{v} \in \mathbb{R}^d$, $c \in \mathbb{R}$, we define an infinite width net with a linear unit by $h_{\alpha,\boldsymbol{v},c}(\boldsymbol{x}) := h_{\alpha,c}(\boldsymbol{x}) + \boldsymbol{v}^\top \boldsymbol{x}$. We prove in Appendix D that $\overline{R}_1(f)$ is equivalent to:

$$\overline{R}_1(f) = \min_{\alpha \in M(\mathbb{S}^{d-1}\times\mathbb{R}), \boldsymbol{v}\in\mathbb{R}^d, c\in R} \|\alpha\|_1 \;\; s.t. \;\; f = h_{\alpha,\boldsymbol{v},c} \tag{13}$$

$$= \|\alpha^*\|_1, \tag{14}$$

where α^* is the unique *even* measure[4] such that $f = h_{\alpha^*,\boldsymbol{v}^*,c^*}$ for some $\boldsymbol{v}^* \in \mathbb{R}^d$ and $c^* \in \mathbb{R}$. The uniqueness of the solution to (13) allows for a much cleaner description of $\overline{R}_1(f)$ in function space relative to $\overline{R}(f)$, as we show in Section 4.

The following lemma, proved in Appendix D, will be used in some our examples, and shows that $\overline{R}_1(f)$ captures the expected representational cost for a ReLU net with finitely many units:

Lemma 1. *Let $f(\boldsymbol{x}) = \sum_{i=1}^k a_i[\boldsymbol{w}_i^\top \boldsymbol{x} - b_i]_+ + \boldsymbol{v}^\top \boldsymbol{x} + c$ where the weight and bias pairs $\{(\boldsymbol{w}_i, b_i)\}_{i=1}^k$ are distinct under the identification $(\boldsymbol{w}, b) = (-\boldsymbol{w}, -b)$. Then $\overline{R}_1(f) = \sum_{i=1}^k |a_i|$.*

3 The Radon transform and its dual

Our characterization of the representational cost $\overline{R}_1(f)$ is posed in terms of *the Radon transform* — a transform that is fundamental to computational imaging, and whose inverse is the basis of image reconstruction in computed tomography. For an investigation of its properties and applications, see Helgason (1999). Here we give a brief review of the Radon transform and its dual as needed for the derivation in Section 4; readers familiar with these topics can skip to Section 4.

The Radon transform $\mathcal{R}$ represents a function $f : \mathbb{R}^d \to \mathbb{R}$ in terms of its integrals over all possible hyperplanes in $\mathbb{R}^d$, as parameterized by the unit normal direction to the hyperplane $\boldsymbol{w} \in \mathbb{S}^{d-1}$ and the signed distance of the hyperplane from the origin $b \in \mathbb{R}$:

$$\mathcal{R}\{f\}(\boldsymbol{w}, b) := \int_{\boldsymbol{w}^\top \boldsymbol{x} = b} f(\boldsymbol{x})\, ds(\boldsymbol{x}) \;\; \text{for all} \;\; (\boldsymbol{w}, b) \in \mathbb{S}^{d-1} \times \mathbb{R} \tag{15}$$

where $ds(\boldsymbol{x})$ represents integration with respect to $(d-1)$-dimensional surface measure on the hyperplane $\boldsymbol{w}^\top \boldsymbol{x} = b$. Note the Radon transform is an *even* function, *i.e.*, $\mathcal{R}\{f\}(\boldsymbol{w}, b) = \mathcal{R}\{f\}(-\boldsymbol{w}, -b)$ for all $(\boldsymbol{w}, b) \in \mathbb{S}^{d-1} \times \mathbb{R}$ since the equations $\boldsymbol{w}^\top \boldsymbol{x} = b$ and $-\boldsymbol{w}^\top \boldsymbol{x} = -b$ determine the same hyperplane.

The Radon transform is invertible for many common spaces of functions, and its inverse is a composition of the *dual Radon transform* $\mathcal{R}^*$ (*i.e.*, the adjoint of $\mathcal{R}$) followed by a filtering step in Fourier domain. The dual Radon transform $\mathcal{R}^*$ maps a function $\varphi : \mathbb{S}^{d-1} \times \mathbb{R} \to \mathbb{R}$ to a function over $\boldsymbol{x} \in \mathbb{R}^d$ by integrating over the subset of coordinates $(\boldsymbol{w}, b) \in \mathbb{S}^{d-1} \times \mathbb{R}$ corresponding to all hyperplanes passing through $\boldsymbol{x} \in \mathbb{R}^d$:

$$\mathcal{R}^*\{\varphi\}(\boldsymbol{x}) := \int_{\mathbb{S}^{d-1}} \varphi(\boldsymbol{w}, \boldsymbol{w}^\top \boldsymbol{x})\, d\boldsymbol{w} \;\; \text{for all} \;\; \boldsymbol{x} \in \mathbb{R}^d \tag{16}$$

where $d\boldsymbol{w}$ represents integration with respect to the surface measure of the unit sphere $\mathbb{S}^{d-1}$. The filtering step is described by a power of the (negative) Laplacian $(-\Delta)^{s/2}$, which for any positive integer $s > 0$ is the operator defined in Fourier domain by

$$\widehat{(-\Delta)^{s/2} f}(\boldsymbol{\xi}) = \|\boldsymbol{\xi}\|^s \widehat{f}(\boldsymbol{\xi}) \tag{17}$$

where $\widehat{g}(\boldsymbol{\xi}) := \int g(\boldsymbol{x}) e^{-i2\pi\xi^\top \boldsymbol{x}} d\boldsymbol{x}$ is the d-dimensional Fourier transform at the Fourier domain (frequency) variable $\boldsymbol{\xi} \in \mathbb{R}^d$. When s is a even integer, $(-\Delta)^{s/2}$ is the same as applying the negative Laplacian $s/2$ times, while if s is odd it is a pseudo-differential operator given by convolution with a singular kernel.

An inversion formula for the Radon transform is given by $f = \gamma_d(-\Delta)^{(d-1)/2}\mathcal{R}^*\{\mathcal{R}\{f\}\}$, where γ_d is a constant depending on dimension d, which holds for f belonging to many common function

[4] Roughly speaking, a measure α is even if $d\alpha(\boldsymbol{w}, b) = d\alpha(-\boldsymbol{w}, -b)$; see Appendix B for a precise definition.

spaces Helgason (1999). The dual Radon transform is also invertible by a similar formula, albeit under more restrictive conditions on the function space. We use the following formula due to Solmon (1987) that holds for all Schwartz class functions[5], which we denote by $\mathcal{S}(\mathbb{S}^{d-1} \times \mathbb{R})$:

Lemma 2 (Solmon (1987))**.** *If φ is an even function*[6], i.e.*, $\varphi(-\boldsymbol{w}, -b) = \varphi(\boldsymbol{w}, b)$ for all $(\boldsymbol{w}, b) \in \mathbb{S}^{d-1} \times \mathbb{R}$, belonging to the Schwartz class $\mathcal{S}(\mathbb{S}^{d-1} \times \mathbb{R})$, then*

$$\gamma_d \mathcal{R}\{(-\Delta)^{(d-1)/2} \mathcal{R}^*\{\varphi\}\} = \varphi, \tag{18}$$

where $\gamma_d = \frac{1}{2(2\pi)^{d-1}}$.

Additional properties of the Radon transform used in this work are given in Appendix A.

4 Representational cost in function space: the $\mathcal{R}$-norm

Our starting point is to relate the Laplacian of an infinite width net to the dual Radon transform of its defining measure. In particular, consider an infinite width net f defined in terms of a smooth density $\alpha(\boldsymbol{w}, b)$ over $\mathbb{S}^{d-1} \times \mathbb{R}$ that decreases rapidly in b, so that we can write

$$f(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1} \times \mathbb{R}} \left([\boldsymbol{w}^\top \boldsymbol{x} - b]_+ - [-b]_+\right) \alpha(\boldsymbol{w}, b)\, d\boldsymbol{w}\, db + \boldsymbol{v}^\top \boldsymbol{x} + c. \tag{19}$$

Differentiating twice inside the integral, the Laplacian $\Delta f(\boldsymbol{x}) = \sum_{i=1}^d \partial_{x_i}^2 f(\boldsymbol{x})$ is given by

$$\Delta f(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1} \times \mathbb{R}} \delta(\boldsymbol{w}^\top \boldsymbol{x} - b) \alpha(\boldsymbol{w}, b)\, d\boldsymbol{w}\, db = \int_{\mathbb{S}^{d-1}} \alpha(\boldsymbol{w}, \boldsymbol{w}^\top \boldsymbol{x})\, d\boldsymbol{w}. \tag{20}$$

where $\delta(\cdot)$ denotes a Dirac delta. We see that the right-hand side of (20) is precisely the dual Radon transform of α, *i.e.*, we have shown $\Delta f = \mathcal{R}^*\{\alpha\}$. Applying the inversion formula (18) to this identity immediately gives the following result:

Lemma 3. *Suppose $f = h_{\alpha, \boldsymbol{v}, c}$ for some $\alpha \in \mathcal{S}\left(\mathbb{S}^{d-1} \times \mathbb{R}\right)$ with α even, and $\boldsymbol{v} \in \mathbb{R}^d$, $c \in \mathbb{R}$. Then $\alpha = -\gamma_d \mathcal{R}\{(-\Delta)^{(d+1)/2} f\}$.*

As a consequence of Lemma 3, and the characterization of $\overline{R}_1(f)$ given in (13), we have the following result, which for simplicity we state in the case d odd[7].

Proposition 1. *Suppose d is odd. If both $f \in L^1(\mathbb{R}^d)$ and $\Delta^{(d+1)/2} f \in L^1(\mathbb{R}^d)$, then*

$$\overline{R}_1(f) = \gamma_d \|\mathcal{R}\{\Delta^{(d+1)/2} f\}\|_1 = \gamma_d \|\partial_b^{d+1} \mathcal{R}\{f\}\|_1 < \infty \tag{21}$$

with $\gamma_d = \frac{1}{2(2\pi)^{d-1}}$.

Here we made use of the intertwining property of the intertwining property of the Laplacian and the Radon transform to write $\mathcal{R}\{\Delta^{(d+1)/2} f\} = \pm \partial_b^{d+1} \mathcal{R}\{f\}$ (see Appendix A).

Proposition 1 suggests that for an *arbitrary* function f we might hope to have $\overline{R}_1(f)$ equal to one of the expressions in (21). However, for many functions of interest these quantities are not well-defined classically. For example, the finite width net $f(\boldsymbol{x}) = \sum_{i=1}^n a_i [\boldsymbol{w}_i^\top \boldsymbol{x} - b_i]_+$ is a piecewise linear function that is non-smooth along each hyperplane $\boldsymbol{w}_i^\top \boldsymbol{x} = b_i$, so its derivatives can only be understood in the sense of generalized functions or distributions. Similarly, in this case the Radon transform of f is not well-defined since f is unbounded and not integrable along hyperplanes.

Instead, we use duality to define a functional (the "$\mathcal{R}$-norm") that extends to the more general case where f is possibly non-smooth or not integrable along hyperplanes. In particular, we define a

[5] *i.e.*, functions φ defined on $\mathbb{S}^{d-1} \times \mathbb{R}$ that are C^∞-smooth such that φ and all its partial derivatives decrease faster than $O(|b|^{-N})$ as $|b| \to \infty$ for any $N \geq 0$

[6] The assumption that φ is even is necessary since odd functions are annihilated by $\mathcal{R}^*$.

[7] For d even an analogous result holds with the pseudo-differential operators $(-\Delta)^{(d+1)/2}$ and $(-\partial_b^2)^{(d+1)/2}$ in place of $\Delta^{(d+1)/2}$ and ∂_b^{d+1}; see Section 3.

functional on the space of all Lipschitz functions[8], which we denote by $\mathrm{Lip}(\mathbb{R}^d)$. The main idea is to re-express the L^1-norm in (21) as a supremum of the inner product over a space of dual functions φ, *i.e.*, using the fact $\mathcal{R}^*$ is the adjoint of $\mathcal{R}$ and the Laplacian Δ is self-adjoint we write

$$\|\mathcal{R}\{\Delta^{(d+1)/2}\}\|_1 = \sup_{\|\varphi\|_\infty \leq 1} \langle \mathcal{R}\{\Delta^{(d+1)/2} f\}, \varphi\rangle = \sup_{\|\varphi\|_\infty \leq 1} \langle f, \Delta^{(d+1)/2}\mathcal{R}^*\{\varphi\}\rangle \tag{22}$$

then restrict φ to a space where $\Delta^{(d+1)/2}\mathcal{R}^*\{\varphi\}$ is always well-defined. This is made formal below.

Definition 1. *For any function $f \in \mathrm{Lip}(\mathbb{R}^d)$ define its $\mathcal{R}$-norm[9], denoted as $\|f\|_{\mathcal{R}}$, by*

$$\|f\|_{\mathcal{R}} := \sup\left\{-\gamma_d\langle f, (-\Delta)^{(d+1)/2}\mathcal{R}^*\{\psi\}\rangle : \psi \in \mathcal{S}(\mathbb{S}^{d-1}\times\mathbb{R}), \psi \textit{ even }, \|\psi\|_\infty \leq 1\right\}. \tag{23}$$

where $\gamma_d = \frac{1}{2(2\pi)^{d-1}}$, $\mathcal{S}(\mathbb{S}^{d-1}\times\mathbb{R})$ is the space of Schwartz functions on $\mathbb{S}^{d-1}\times\mathbb{R}$, and $\langle f, g\rangle := \int_{\mathbb{R}^d} f(\boldsymbol{x})g(\boldsymbol{x})d\boldsymbol{x}$. If $f \notin \mathrm{Lip}(\mathbb{R}^d)$ we define $\|f\|_{\mathcal{R}} = +\infty$.

We prove in Appendix that the $\mathcal{R}$-norm is well-defined, though not always finite, for all Lipschitz functions and, whether finite or infinite, is always equal to the representation cost:

Theorem 1. *$\overline{R}_1(f) = \|f\|_{\mathcal{R}}$ for all functions f. In particular, $\overline{R}_1(f)$ is finite if and only if f is Lipschitz and $\|f\|_{\mathcal{R}}$ is finite.*

We give the proof of Theorem 1 in the Appendix C, but the following example illustrates many key elements of the proof.

Example 1. *We show $\|f\|_{\mathcal{R}}$ is consistent with $\overline{R}_1(f)$ in the case where f is a finite-width two-layer ReLU network. First, consider the case where f consists of a single ReLU unit $f(\boldsymbol{x}) = a[\boldsymbol{w}^\top\boldsymbol{x}-b]_+$. Then by Lemma 1, $\overline{R}_1(f) = |a|$. Note that $\Delta f(\boldsymbol{x}) = a\,\delta(\boldsymbol{w}^\top\boldsymbol{x}-b)$ in a distributional sense,* i.e., *for any smooth test function φ we have $\langle\Delta f, \varphi\rangle = \langle f, \Delta\varphi\rangle = a\int\varphi(\boldsymbol{x})\delta(\boldsymbol{w}^\top\boldsymbol{x}-b)d\boldsymbol{x} = a\mathcal{R}\{\varphi\}(\boldsymbol{w}, b)$. So for any even $\psi \in \mathcal{S}(\mathbb{S}^{d-1}\times\mathbb{R})$ we have*

$$-\gamma_d\langle f, (-\Delta)^{(d+1)/2}\mathcal{R}^*\{\psi\}\rangle = \gamma_d\langle\Delta f, (-\Delta)^{(d-1)/2}\mathcal{R}^*\{\psi\}\rangle \tag{24}$$

$$= a\,\gamma_d\mathcal{R}\{(-\Delta)^{(d-1)/2}\mathcal{R}^*\{\psi\}\}(\boldsymbol{w}, b) \tag{25}$$

$$= a\,\psi(\boldsymbol{w}, b) \tag{26}$$

where in the last step we used the inversion formula (18). *Since the supremum defining $\|f\|_{\mathcal{R}}$ is over all even $\psi \in \mathcal{S}(\mathbb{S}^{d-1}\times\mathbb{R})$ such that $\|\psi\|_\infty \leq 1$, taking any ψ^* such that $\psi^*(\boldsymbol{w}, b) = sign(a)$ and $\|\psi^*\|_\infty = 1$ we see that $\|f\|_{\mathcal{R}} = \overline{R}(f) = |a|$ as expected. The general case follows by linearity: let $g(\boldsymbol{x}) = \sum_{i=1}^k a_i[\boldsymbol{w}_i^\top\boldsymbol{x} - b_i]_+$ with all pairs $\{(\boldsymbol{w}_i, b_i)\}_{i=1}^k$ distinct. Then for any $\psi \in \mathcal{S}(\mathbb{S}^{d-1}\times\mathbb{R})$ we can show*

$$-\gamma_d\langle g, (-\Delta)^{(d+1)/2}\mathcal{R}^*\{\psi\}\rangle = \sum_{i=1}^k a_i\,\psi(\boldsymbol{w}_i, b_i). \tag{27}$$

Letting ψ^* be any even Schwartz function such that $\psi^*(\boldsymbol{w}_i, b_i) = sign(a_i)$ for all $i = 1, ..., k$ and $\|\psi^*\|_\infty = 1$, we see that $\|g\|_{\mathcal{R}} = \sum_{i=1}^k |a_i| = \overline{R}_1(g)$, as expected.

The representational cost $\overline{R}(f)$ defined without the unregularized linear unit is more difficult to explicitly characterize than $\overline{R}_1(f)$. However, we prove that $\overline{R}(f)$ is finite if and only if $\overline{R}_1(f)$ is finite, and give bounds for $\overline{R}(f)$ in terms of $\|f\|_{\mathcal{R}}$ and the norm of the gradient of the function "at infinity", similar to the expressions derived in Savarese et al. (2019) in the 1D setting.

Theorem 2. *$\overline{R}(f)$ is finite if and only if $\|f\|_{\mathcal{R}}$ is finite, in which case we have the bounds*

$$\max\{\|f\|_{\mathcal{R}}, 2\|\nabla f(\infty)\|\} \leq \overline{R}(f) \leq \|f\|_{\mathcal{R}} + 2\|\nabla f(\infty)\| \tag{28}$$

where $\nabla f(\infty) := \lim_{r\to\infty}\frac{1}{c_d r^{d-1}}\oint_{\|\boldsymbol{x}\|=r}\nabla f(\boldsymbol{x})ds(\boldsymbol{x}) \in \mathbb{R}^d$. In particular, if $\nabla f(\infty) = \boldsymbol{0}$ then $\overline{R}(f) = \overline{R}_1(f) = \|f\|_{\mathcal{R}}$.

[8] Recall that f is Lipschitz if there exists a constant L (depending on f) such that $|f(\boldsymbol{x})-f(\boldsymbol{y})| \leq L\|\boldsymbol{x}-\boldsymbol{y}\|$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$.

[9] Strictly speaking, the functional $\|\cdot\|_{\mathcal{R}}$ is not a norm, but it is a semi-norm on the space of functions for which it is finite; see Appendix F

We give the proof of Theorem 2 in Appendix E. The lower bound $\max\{\|f\|_{\mathcal{R}}, 2\|\nabla f(\infty)\|\}$ is analogous to the expression for the 1D representational cost (1) obtained in Savarese et al. (2019). From this, one might speculate that $\overline{R}(f)$ is equal to $\max\{\|f\|_{\mathcal{R}}, 2\|\nabla f(\infty)\|_2\}$. However, in Appendix E we show this is not the case: there are examples of functions f in all dimensions such that $\overline{R}(f)$ attains the upper bound in a non-trivial way (*e.g.*, $f(x,y) = |x| + y$ in $d = 2$).

4.1 Properties of the $\mathcal{R}$-norm

In Appendix F we prove several useful properties for the $\mathcal{R}$-norm. In particular, we show the $\mathcal{R}$-norm is in fact a *semi-norm*, *i.e.*, is absolutely homogeneous and satisfies the triangle inequality, while $\|f\|_{\mathcal{R}} = 0$ if and only if f is affine. We also show $\mathcal{R}$-norm is invariant to coordinate translation and rotations, and prove the following scaling law under contractions/dilation:

Proposition 2. *If $f_\varepsilon(\boldsymbol{x}) := f(\boldsymbol{x}/\varepsilon)$ for any $\varepsilon > 0$, then $\|f_\varepsilon\|_{\mathcal{R}} = \varepsilon^{-1}\|f\|_{\mathcal{R}}$*

Proposition 2 shows that "spikey" functions will necessarily have large $\mathcal{R}$-norm. For example, let f be any non-negative function supported on the ball of radius 1 with maximum height 1 such that $\|f\|_{\mathcal{R}}$ is finite. Then the contraction f_ε is supported on the ball of radius ε with maximum height 1, but $\|f_\varepsilon\|_{\mathcal{R}} = \varepsilon^{-1}\|f\|_{\mathcal{R}}$ blows up as $\varepsilon \to 0$.

From a generalization perspective, the fact that the $\mathcal{R}$-norm blows up with contractions is a desirable property, since otherwise the minimum norm fit to data would be spikes on data points. In particular, this is what would happen if the representational cost involved derivatives lower than $d+1$, and so in this sense it is not a coincidence that $\|f\|_{\mathcal{R}}$ involves derivatives of order $d+1$.

Finally, we show the smoothness requirements of the $\mathcal{R}$-norm are also reflected in Fourier domain. In particular, we show that for a broad class of functions in order $\mathcal{R}$-norm to be finite the Fourier transform of f must decay rapidly along every ray. A precise statement is given in Proposition 13 in Appendix F.

5 Consequences, Applications and Discussion

Our characterization of the representational cost for multivariate functions in terms of the $\mathcal{R}$-norm is unfortunately not as simple as the characterization in the univariate case. Nevertheless, it is often easy to evaluate, and is a powerful tool for studying the representational power of bounded norm ReLU networks.

5.1 Sobolev spaces

Here we relate Sobolev spaces and the $\mathcal{R}$-norm. The key result is the following upper bound, which is proved in Appendix G.

Proposition 3. *If $f \in \mathrm{Lip}(\mathbb{R}^d)$ and $(-\Delta)^{(d+1)/2} f$ exists in a weak sense then*

$$\|f\|_{\mathcal{R}} \leq c_d \gamma_d \|(-\Delta)^{(d+1)/2} f\|_1. \tag{29}$$

where $c_d = \int_{\mathbb{S}^{d-1}} d\boldsymbol{w} = \frac{2\pi^{d/2}}{\Gamma(d/2)}$, and $\gamma_d = \frac{1}{2(2\pi)^{d-1}}$.

Recall that if the dimension d is odd then $(-\Delta)^{(d+1)/2}$ is just an integer power of the negative Laplacian, which is a linear combination of partial derivatives of order $d+1$. Hence, we have $\|(-\Delta)^{(d+1)/2} f\|_1 \leq c_d \gamma_d \|f\|_{W^{d+1,1}}$, where $\|f\|_{W^{d+1,1}}$ is the Sobolev norm given by the sum of L^1-norm of f and the L^1-norms of all its weak derivatives up to order $d+1$. This gives the following immediate corollary to Proposition 3:

Corollary 1. *Suppose d is odd. If f belongs to the Sobolev space $W^{d+1,1}(\mathbb{R}^d)$,* i.e.*, f and all its weak derivatives up to order $d+1$ are in $L^1(\mathbb{R}^d)$, then $\|f\|_{\mathcal{R}}$ is finite and $\|f\|_{\mathcal{R}} \leq c_d \gamma_d \|f\|_{W^{d+1,1}}$.*

Corollary 1 shows that the space of functions with finite $\mathcal{R}$-norm is "dense" in the space of all functions, in the sense that it contains a full Sobolev space.

5.2 Radial bump functions

Here we study the case where f is a radially symmetric function, *i.e.*, $f(\boldsymbol{x}) = g(\|\boldsymbol{x}\|)$ for some function $g : [0, \infty) \to \mathbb{R}$. In this case, the $\mathcal{R}$-norm is expressible entirely in terms of derivatives of the radial profile function g, as shown in the following result, which is proved in Appendix H.

Proposition 4. *Suppose $d \geq 3$ is odd. If $f \in L^1(\mathbb{R}^d)$ with $f(\boldsymbol{x}) = g(\|\boldsymbol{x}\|)$ then*

$$\|f\|_{\mathcal{R}} = \frac{2}{(d-2)!} \int_0^\infty \left| \partial^{(d+1)} \rho(b) \right| db. \tag{30}$$

where $\rho(b) = \int_b^\infty g(t)(t^2 - b^2)^{(d-3)/2} t \, dt$,

For example, in the $d = 3$ dimensional case, we have

$$\|f\|_{\mathcal{R}} = 2 \int_0^\infty |b\, \partial^3 g(b) + 3\partial^2 g(b)| db, \quad (d = 3) \tag{31}$$

More generally, for any odd dimension $d \geq 3$ a simple induction shows (30) is equivalent to

$$\|f\|_{\mathcal{R}} = \frac{2}{(d-2)!} \int_0^\infty |Q_d\{g\}(b)| db \tag{32}$$

where Q_d is a differential operator of degree $(d+3)/2$ having the form $Q_d = \sum_{k=2}^{(d+3)/2} p_{k,d}(b) \partial^k$ where each $p_{k,d}(b)$ is a polynomial in b of degree $k-2$. In particular, if the weak derivative $\partial^{(d+1)/2} g$ exists and has bounded variation, then $\|f\|_{\mathcal{R}}$ is finite.

Example 2. *Consider the radial bump function $f(\boldsymbol{x}) = g(\|\boldsymbol{x}\|)$ with $\boldsymbol{x} \in \mathbb{R}^3$ where*

$$g(r) = \begin{cases} (1-r^2)^2 & \textit{if } 0 \leq r < 1 \\ 0 & \textit{if } r \geq 1. \end{cases} \tag{33}$$

which is non-negative, supported on the unit ball, and has maximum height $f(\mathbf{0}) = 1$, and let $f_\varepsilon(\boldsymbol{x}) = f(\boldsymbol{x}/\varepsilon)$ be the contraction of f to a ball of radius ε with the same height. Then using formula (31)*, and the dilation property* (2)*, we can compute*

$$\|f_\varepsilon\|_{\mathcal{R}} = \|f\|_{\mathcal{R}} / \varepsilon = 16(1 + \tfrac{1}{5}(5 + 2\sqrt{5}))/\varepsilon. \tag{34}$$

Note that if we move up to dimension $d = 5$, then the function defined by (33) no longer has finite norm since its derivatives of order $(d+3)/2 = 4$ do not exist; this phenomenon is explored in more detail in the next example.

Example 3. *Suppose $d \geq 3$ is odd. Consider the radial bump function $f_{d,k}(\boldsymbol{x}) = g_{d,k}(\|\boldsymbol{x}\|)$ with $\boldsymbol{x} \in \mathbb{R}^d$ where*

$$g_{d,k}(r) = \begin{cases} (1-r^2)^k & \textit{if } 0 \leq r < 1 \\ 0 & \textit{if } r \geq 1. \end{cases} \tag{35}$$

for any $k > 0$. Then we prove $\|f_{d,k}\|_{\mathcal{R}}$ is finite if and only if $k \geq \frac{d+1}{2}$. To illustrate the scaling with dimension d, we set $k_d = (d+1)/2 + 2$. In Appendix H we prove the bounds $(d+5)d \leq \|f_{d,k_d}\|_{\mathcal{R}} \leq 2d(d+5)$, hence we have the scaling $\|f_{d,k_d}\|_{\mathcal{R}} \sim d^2$. Similarly, a contraction of f_{d,k_d} to the ball of radius ε will have $\mathcal{R}$-norm scaling as $\sim d^2/\varepsilon$

The next example[10] shows there there is a universal choice of radial bump function in all (odd) dimensions with finite $\mathcal{R}$-norm:

Example 4. *Suppose $d \geq 3$ is odd. Consider the radial bump function $f(\boldsymbol{x}) = g(\|\boldsymbol{x}\|)$ with $\boldsymbol{x} \in \mathbb{R}^d$ where*

$$g(r) = \begin{cases} e^{-\frac{1}{1-r^2}} & \textit{if } 0 \leq r < 1 \\ 0 & \textit{if } r \geq 1. \end{cases} \tag{36}$$

Since g is C^∞-smooth and its derivatives of all orders are L^1-bounded, f has finite $\mathcal{R}$-norm by Proposition 4.

[10] The existence of such a radial function was noted in parallel work by Matus Telgarsky. Discussions with Telgarsky motivated us to construct and analyze it using the $\mathcal{R}$-norm.

5.3 Piecewise Linear functions

Every finite-width two-layer ReLU network is a continuous piecewise linear function. However, the reverse implication is not true. For example, in dimensions two and above no compactly supported piecewise linear function is expressible as a finite-width two-layer ReLU network. A natural question then is: what piecewise linear functions are represented by bounded norm infinite-width nets? In particular, can a compactly supported piecewise linear function be represented by a bounded norm infinite-width net? The following result, proved in Appendix F, shows this is not the case:

Proposition 5. *A continuous piecewise linear function with compact support has infinite $\mathcal{R}$-norm.*

This result suggests that the space of piecewise linear functions expressible as a bounded norm infinite-width two-layer ReLU net is not qualitatively different than those captured by finite-width nets. We go further and make the following conjecture:

Conjecture 1. *A continuous piecewise linear function f (with finitely many pieces) has finite $\mathcal{R}$-norm if and only if it is a two-layer ReLU network with finitely many units.*

5.4 Depth Separation

In an effort to understand the power of deeper networks, there has been much work showing how some functions can be much more easily approximated *in terms of number of required units* by deeper networks compared to shallower ones, including results showing how functions that can be well-approximated by three-layer networks require a much larger number of units to approximate if using a two-layer network (*e.g.*, Pinkus (1999); Telgarsky (2016); Liang & Srikant (2016); Safran & Shamir (2017); Yarotsky (2017)). The following example shows that, also in terms of the norm, such a depth separation exists for ReLU nets:

Example 5. *The pyramid function $f(\boldsymbol{x}) = [1 - \|\boldsymbol{x}\|_1]_+$ has infinite norm as a two-layer ReLU network ($\overline{R}(f) = +\infty$), but can be exactly represented as a finite-width three-layer ReLU network.*

Interestingly, this result shows that, in terms of the norm, we have a qualitive rather then quantitative depth separation: the required norm with three layers is finite, while with only two layers it is not merely very large, but *infinite*. In contrast, in standard depth separation results, the separation is quantitative: we can compensate for a decrease in depth and use more neurons to achieve the same approximation quality. It would be interesting to further strengthen Example 5 by obtaining a quantitative lower bound on the norm required to ϵ-approximate the pyramid with an infinite-width two-layer network.

5.5 The $\mathcal{R}$-norm is not a RKHS norm

There is an ongoing debate in the community on whether neural network learning can be simulated or replicated by kernel machines with the "right" kernel. In this context, it is interesting to ask whether the inductive bias we uncover can be captured by a kernel, or in other words whether the $\mathcal{R}$-norm is an RKHS (semi) norm. The answer is no:

Proposition 6. *The $\mathcal{R}$-norm is not a RKHS (semi-)norm.*

This is seen immediately by the failure of the parallelogram law to hold. For example, if $f_1(\boldsymbol{x}) = [\boldsymbol{w}_1^\top \boldsymbol{x}]_+$, $f_2 = [\boldsymbol{w}_2^\top \boldsymbol{x}]_+$ with $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{S}^{d-1}$ distinct, then by Lemma 1 we have $\|f_1\|_{\mathcal{R}} = \|f_2\|_{\mathcal{R}} = 1$, while $\|f_1 + f_2\|_{\mathcal{R}} = \|f_1 - f_2\|_{\mathcal{R}} = 2$, and so $2(\|f_1\|_{\mathcal{R}}^2 + \|f_2\|_{\mathcal{R}}^2) \neq \|f_1 + f_2\|_{\mathcal{R}}^2 + \|f_1 - f_2\|_{\mathcal{R}}^2$.

5.6 Generalization implications

Neyshabur et al. (2015) shows that training an unbounded size neural net while regularizing the ℓ_2 norm of the weights results in a sample complexity proportional to a variant[11] of $\overline{R}(f)$. This paper gives an explicit characterization of $\overline{R}(f)$ and thus of the sample complexity of learning a function using regularized unbounded-width neural networks.

[11] Their analysis does not allow for unregularized bias, but can be extended to allow for it.

REFERENCES

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.

Andrew R Barron. Approximation and estimation bounds for artificial neural networks. *Machine learning*, 14(1):115–133, 1994.

Peter L Bartlett. For valid generalization the size of the weights is more important than the size of the network. In *Neural Information Processing Systems (NeurIPS)*, pp. 134–140, 1997.

Yoshua Bengio, Nicolas L Roux, Pascal Vincent, Olivier Delalleau, and Patrice Marcotte. Convex neural networks. In *Neural Information Processing Systems (NeurIPS)*, pp. 123–130, 2006.

Jan Boman and Filip Lindskog. Support theorems for the Radon transform and Cramér-Wold theorems. *Journal of theoretical probability*, 22(3):683–710, 2009.

Emmanuel J Candès. Harmonic analysis of neural networks. *Applied and Computational Harmonic Analysis*, 6(2):197–218, 1999.

Emmanuel J Candès and David L Donoho. Ridgelets: A key to higher-dimensional intermittency? *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 357(1760):2495–2509, 1999.

Sean M. Carroll and Bradley W. Dickinson. Construction of neural nets using the Radon transform. In *International Joint Conference on Neural Networks*, volume 1, pp. 607–611, 1989. doi: 10.1109/IJCNN.1989.118639.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 9461–9471. Curran Associates, Inc., 2018. URL `http://papers.nips.cc/paper/8156-implicit-bias-of-gradient-descent-on-linear-convolutional-networks.pdf`.

Sigurdur Helgason. *The Radon transform*. Springer, 1999.

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

Yoshifusa Ito. Representation of functions by superpositions of a step or sigmoid function and their applications to neural network theory. *Neural Networks*, 4(3):385–394, 1991.

Shiyu Liang and R. Srikant. Why Deep Neural Networks for Function Approximation? In *ICLR*, 2016. URL `http://arxiv.org/abs/1610.04161`.

Paul Malliavin. *Integration and probability*, volume 157. Springer Science & Business Media, 2012.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory (COLT)*, pp. 1376–1401, 2015.

Behnam Neyshabur, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*, 2017.

Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta numerica*, 8: 143–195, 1999.

Itay Safran and Ohad Shamir. Depth-Width Tradeoffs in Approximating Natural Functions with Neural Networks. *ICML*, pp. 1–27, 2017. ISSN 1938-7228. URL `http://arxiv.org/abs/1610.09887`.

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 (COLT)*, 2019.

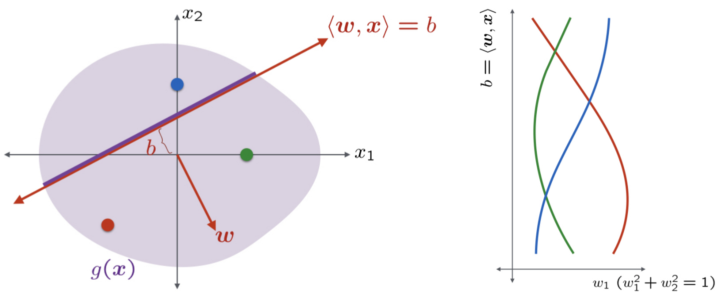
Donald C Solmon. Asymptotic formulas for the dual Radon transform and applications. *Mathematische Zeitschrift*, 195(3):321–343, 1987.

Sho Sonoda and Noboru Murata. Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis*, 43(2):233–268, 2017.

Matus Telgarsky. Benefits of depth in neural networks. *COLT*, Feb 2016. URL `http://arxiv.org/abs/1602.04485`.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017. ISSN 18792782. doi: 10.1016/j.neunet.2017.07.002.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.


x_2
$\langle \boldsymbol{w}, \boldsymbol{x} \rangle = b$
x_1
b
$\boldsymbol{w}$
$g(\boldsymbol{x})$
$b = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$
$w_1 \; (w_1^2 + w_2^2 = 1)$

(a) Radon transform

(b) Dual Radon transform

Figure 1: The Radon transform and its dual. (a) Illustration of the Radon transform in (15) for dimension $d = 2$. The red line of points $\boldsymbol{x}$ satisfies $\boldsymbol{w}^\top \boldsymbol{x} = b$; $\boldsymbol{w}$ determines the line orientation (angle relative to the coordinate axes) and b determines its offset from the origin. (b) Illustration of the support of the Radon transform of Dirac deltas centered at the colored points. If a function f is a superposition of such delta functions, then $\mathcal{R}\{f\}$ is the sum of the curves in (b); this is typically referred to as a "sinogram". Furthermore, the dual Radon transform $\mathcal{R}\{\varphi\}$ given in equation (16) integrates a function $\varphi(\boldsymbol{w}, b)$ over a curve like one of the three in (b).

Appendices

A Additional properties of the Radon Transform

Figure 1 illustrates the Radon transform and its dual in dimension $d = 2$.

We will often use the fact that the Radon transform is a bounded linear operator from $L^1(\mathbb{R}^d)$ to $L^1(\mathbb{S}^{d-1} \times \mathbb{R})$, *i.e.*, if $f \in L^1(\mathbb{R}^d)$ then $\mathcal{R}\{f\} \in L^1(\mathbb{R}^d)$. In particular, if $f \in L^1(\mathbb{R}^d)$ then $\mathcal{R}\{f\}$ is defined almost everywhere on $\mathbb{S}^{d-1} \times \mathbb{R}$, and the function $\mathcal{R}\{f\}(\boldsymbol{w}, \cdot)$ is in $L^1(\mathbb{R})$ for all $\boldsymbol{w} \in \mathbb{S}^{d-1}$.

Here we recall the *Fourier slice theorem* for Radon transform (see, *e.g.*, Helgason (1999)): Let $f \in L^1(\mathbb{R}^d)$, then for all $\sigma \in \mathbb{R}$ and $\boldsymbol{w} \in \mathbb{S}^{d-1}$ we have

$$\mathcal{F}_b \mathcal{R}\{f\}(\boldsymbol{w}, \sigma) = \widehat{f}(\sigma \cdot \boldsymbol{w}) \tag{37}$$

where $\mathcal{F}_b$ indicates the 1-D Fourier transform in the offset variable b. From this it is easy to establish the following *intertwining property* of the Laplacian and the Radon Transform: assuming f and Δf are in $L^1(\mathbb{R}^d)$, we have

$$\mathcal{R}\{\Delta f\} = \partial_b^2 \mathcal{R}\{f\} \tag{38}$$

where ∂_b is the partial derivative in the offset variable b. More generally, for any positive integer s we have

$$\mathcal{R}\{(-\Delta)^{s/2} f\} = (-\partial_b^2)^{s/2} \mathcal{R}\{f\} \tag{39}$$

where fractical powers of $-\partial_b^2$ can be defined in Fourier domain, same as fractional powers of the Laplacian. In particular, for s an even integer $(-\partial_b^2)^{s/2} = (-1)^{s/2} \partial_b^s$ while for s an odd integer, $(-\partial_b^2)^{s/2} = (\mathcal{H}\partial_b)^s$ where $\mathcal{H}$ is the Hilbert transform in the offset variable b.

B Infinite-width Nets

Measures and infinite-width nets Let α be a signed measure [12] defined on $\mathbb{S}^{d-1} \times \mathbb{R}$, and let $\|\alpha\|_1 = \int d|\alpha|$ denote its total variation norm. We let $M(\mathbb{S}^{d-1} \times \mathbb{R})$ denote the space of measures α with finite total variation norm. Since $\mathbb{S}^{d-1} \times \mathbb{R}$ is a locally compact space, $M(\mathbb{S}^{d-1} \times \mathbb{R})$ is the Banach space dual of $C_0(\mathbb{S}^{d-1} \times \mathbb{R})$, the space of continuous functions on $\mathbb{S}^{d-1} \times \mathbb{R}$ vanishing at infinity (Malliavin, 2012, Chapter 2, Theorem 6.6), and

$$\|\alpha\|_1 = \sup \left\{ \int \varphi \, d\alpha : \varphi \in C_0(\mathbb{S}^{d-1} \times \mathbb{R}), \|\varphi\|_\infty \leq 1 \right\}. \tag{40}$$

[12] To be precise, we assume α is a signed *Radon* measure; see, *e.g.*, Malliavin (2012) for a formal definition. We omit the word "Radon" and simply call α a measure to avoid confusion with the Radon transform, which is central to this work.

For $\alpha \in M(\mathbb{S}^{d-1} \times \mathbb{R})$ and $\varphi \in C_0(\mathbb{S}^{d-1} \times \mathbb{R})$, we often write $\langle \alpha, \varphi \rangle = \int \varphi d\alpha$.

Note that any $\alpha \in M(\mathbb{S}^{d-1} \times \mathbb{R})$ can be extended uniquely as a linear functional to $C_b(\mathbb{S}^{d-1} \times \mathbb{R})$, the space continuous and bounded functions on $\mathbb{S}^{d-1} \times \mathbb{R}$. In particular, since the function $\varphi(\boldsymbol{w}, b) = [\boldsymbol{w}^\top \boldsymbol{x} - b]_+ - [-b]_+$ belongs to $C_b(\mathbb{S}^{d-1} \times \mathbb{R})$, we see that the infinite width net

$$h_\alpha(\boldsymbol{x}) := \int_{\mathbb{S}^{d-1} \times \mathbb{R}} ([\boldsymbol{w}^\top \boldsymbol{x} - b]_+ - [-b]_+) d\alpha(\boldsymbol{w}, b) \tag{41}$$

is well-defined for all $\boldsymbol{x} \in \mathbb{R}^d$.

Remark 1. Our definition of an infinite-width net in differs slightly from Savarese et al. (2019): we integrate a constant shift of the ReLU $[\boldsymbol{w}^\top \boldsymbol{x} - b]_+ - [-b]_+$ with respect to the measure $\alpha(\boldsymbol{w}, b)$ rather than $[\boldsymbol{w}^\top \boldsymbol{x} - b]_+$ as in Savarese et al. (2019). As shown above this ensures the integral is always well-defined for any measure α with finite total variation. Alternatively, we could have restricted to measures that have finite first moment, *i.e.*, $\int_{\mathbb{S}^{d-1} \times \mathbb{R}} |b| \, d|\alpha|(\boldsymbol{w}, b) < \infty$, which ensures $\widetilde{h}_\alpha(\boldsymbol{x}) := \int_{\mathbb{S}^{d-1} \times \mathbb{R}} [\boldsymbol{w}^\top \boldsymbol{x} - b]_+ d\alpha(\boldsymbol{w}, b)$, as given in Savarese et al. (2019), is always well-defined. However, restricting to measures with finite first moment complicates the function space description, and excludes from our analysis certain functions that are still naturally defined as limits of bounded norm finite width networks, and so we opt for the definition above instead. In the case that α has a finite first moment the difference between definitions is immaterial since h_α and $\widetilde{h}_\alpha$ are equal up to an additive constant, which implies they have the same representational cost under $\overline{R}(\cdot)$ and $\overline{R}_1(\cdot)$.

Even and odd measures We say $\alpha \in M(\mathbb{S}^{d-1} \times \mathbb{R})$ is *even* if

$$\int_{\mathbb{S}^{d-1} \times \mathbb{R}} \varphi(\boldsymbol{w}, b) d\alpha(\boldsymbol{w}, b) = \int_{\mathbb{S}^{d-1} \times \mathbb{R}} \varphi(-\boldsymbol{w}, -b) d\alpha(\boldsymbol{w}, b) \text{ for all } \varphi \in C_0(\mathbb{S}^{d-1} \times \mathbb{R}) \tag{42}$$

or α is *odd* if

$$\int_{\mathbb{S}^{d-1} \times \mathbb{R}} \varphi(\boldsymbol{w}, b) d\alpha(\boldsymbol{w}, b) = -\int_{\mathbb{S}^{d-1} \times \mathbb{R}} \varphi(-\boldsymbol{w}, -b) d\alpha(\boldsymbol{w}, b) \text{ for all } \varphi \in C_0(\mathbb{S}^{d-1} \times \mathbb{R}). \tag{43}$$

It is easy to show every measure $\alpha \in M(\mathbb{S}^{d-1} \times \mathbb{R})$ is uniquely decomposable as $\alpha = \alpha_+ + \alpha_-$ where α_+ is even and α_- is odd. For example, if α has a density $\mu(\boldsymbol{w}, b)$ then α_+ is the measure with density $\mu_+(\boldsymbol{w}, b) = \frac{1}{2}(\mu(\boldsymbol{w}, b) + \mu(-\boldsymbol{w}, -b))$ and α_- is the measure with density $\mu_-(\boldsymbol{w}, b) = \frac{1}{2}(\mu(\boldsymbol{w}, b) - \mu(-\boldsymbol{w}, -b))$.

We let $M(\mathbb{P}^d)$ denote the subspace of all even measures $\alpha \in M(\mathbb{S}^{d-1} \times \mathbb{R})$, and identify its Banach space dual with $C_0(\mathbb{P}^d)$, the subspace of all even functions $\varphi \in C_0(\mathbb{S}^{d-1} \times \mathbb{R})$. Even measures play an important role in our results because of the following observations.

Let $\alpha \in M(\mathbb{S}^{d-1} \times \mathbb{R})$. If $\alpha = \alpha_+ + \alpha_-$ where α_+ is even and α_- is odd, then we have $h_\alpha = h_{\alpha_+} + h_{\alpha_-}$. By the identity $[t]_+ + [-t]_+ = |t|$ we can show

$$h_{\alpha_+}(\boldsymbol{x}) = \frac{1}{2} \int_{\mathbb{S}^{d-1} \times \mathbb{R}} (|\boldsymbol{w}^\top \boldsymbol{x} + b| - |b|) d\alpha_+(\boldsymbol{w}, b). \tag{44}$$

Likewise, by the identity $[t]_+ - [-t]_+ = t$ we have

$$h_{\alpha_-}(\boldsymbol{x}) = \boldsymbol{v}_0^\top \boldsymbol{x}. \tag{45}$$

where $\boldsymbol{v}_0 = \frac{1}{2} \int_{\mathbb{S}^{d-1} \times \mathbb{R}} \boldsymbol{w} d\alpha_-(\boldsymbol{w}, b)$. Hence, h_α decomposes into a sum of a component with absolute value activations and a linear function. In particular, if $f = h_{\alpha, \boldsymbol{v}, c}$ for some $\alpha \in M(\mathbb{S}^{d-1} \times \mathbb{R})$, $\boldsymbol{v} \in \mathbb{R}^d$, $c \in \mathbb{R}$, we can always write $f = h_{\alpha_+, \boldsymbol{v}', c}$ for some other $\boldsymbol{v}' \in \mathbb{R}^d$ where α_+ is the even part of α. In other words, we lose no generality by restricting ourselves to infinite width nets of the form $f = h_{\alpha, \boldsymbol{v}, c}$ where $\alpha \in M(\mathbb{P}^d)$ (*i.e.*, α is even).

We will need the following fact about even and odd decompositions of measures under the total variation norm:

Proposition 7. *Let $\alpha \in M(\mathbb{S}^{d-1} \times \mathbb{R})$ with $\alpha = \alpha_+ + \alpha_-$ where α_+ is even and α_- is odd. Then $\|\alpha_+\|_1 \leq \|\alpha\|_1$ and $\|\alpha_-\|_1 \leq \|\alpha\|_1$.*

Proof. For any $\varphi \in C_0(\mathbb{S}^{d-1} \times \mathbb{R})$ we can write $\varphi = \varphi_+ + \varphi_-$ where $\varphi_+(\boldsymbol{w}, b) = \frac{1}{2}(\varphi(\boldsymbol{w}, b) + \varphi(-\boldsymbol{w}, -b))$ is even and $\varphi_-(\boldsymbol{w}, b) = \frac{1}{2}(\varphi(\boldsymbol{w}, b) - \varphi(-\boldsymbol{w}, -b))$ is odd. Note that $\int \varphi \, d\alpha_+ = \int \varphi_+ \, d\alpha_+$ since $\int \varphi_- d\alpha_+ = 0$. Furthermore, if $|\varphi(\boldsymbol{w}, b)| \leq 1$ for all $(\boldsymbol{w}, b) \in \mathbb{S}^{d-1} \times \mathbb{R}$ we see that $|\varphi_+(\boldsymbol{w}, b)| \leq \frac{1}{2}(|\varphi(\boldsymbol{w}, b)| + |\varphi(-\boldsymbol{w}, -b)|) \leq 1$ for all $(\boldsymbol{w}, b) \in \mathbb{S}^{d-1} \times \mathbb{R}$. Therefore, in the dual definition of $\|\alpha_+\|_1$ given in (40) it suffices to take the supremum over all even functions $\varphi \in C_0(\mathbb{S}^{d-1} \times \mathbb{R})$. Hence,

$$\|\alpha\|_1 = \sup \left\{ \int \varphi \, d\alpha : \varphi \in C_0(\mathbb{S}^{d-1} \times \mathbb{R}), \|\varphi\|_\infty \leq 1 \right\} \tag{46}$$

$$= \sup \left\{ \int \varphi \, d\alpha_+ + \int \varphi \, d\alpha_- : \varphi \in C_0(\mathbb{S}^{d-1} \times \mathbb{R}), \|\varphi\|_\infty \leq 1 \right\} \tag{47}$$

$$\geq \sup \left\{ \int \varphi \, d\alpha_+ + \int \varphi \, d\alpha_- : \varphi \in C_0(\mathbb{S}^{d-1} \times \mathbb{R}), \|\varphi\|_\infty \leq 1, \varphi \text{ even} \right\} \tag{48}$$

$$= \sup \left\{ \int \varphi \, d\alpha_+ : \varphi \in C_0(\mathbb{S}^{d-1} \times \mathbb{R}), \|\varphi\|_\infty \leq 1, \varphi \text{ even} \right\} \tag{49}$$

$$= \|\alpha_+\|_1 \tag{50}$$

A similar argument shows $\|\alpha_-\|_1 \leq \|\alpha\|_1$. □

Lipschitz continuity of infinite-width nets Recall that we define $\mathrm{Lip}(\mathbb{R}^d)$ to be the space of all Lipschitz continuous functions on $\mathbb{R}^d$. For any $f \in \mathrm{Lip}(\mathbb{R}^d)$, define $\|f\|_L := \sup_{\boldsymbol{x} \neq \boldsymbol{y}} \frac{|f(\boldsymbol{x}) - f(\boldsymbol{y})|}{\|\boldsymbol{x} - \boldsymbol{y}\|}$, *i.e.*, the smallest possible Lipschitz constant. The following result shows that $\mathrm{Lip}(\mathbb{R}^d)$ is a natural space to work in when considering infinite-width nets:

Proposition 8. *Let $f = h_{\alpha, \boldsymbol{v}, c}$ for any $\alpha \in M(\mathbb{S}^{d-1} \times \mathbb{R}), \boldsymbol{v} \in \mathbb{R}^d, c \in \mathbb{R}$. Then $f \in \mathrm{Lip}(\mathbb{R}^d)$ with $\|f\|_L \leq \|\alpha\|_1 + \|\boldsymbol{v}\|$.*

Proof. First we prove for all even $\alpha \in M(\mathbb{P}^d)$, $\|h_\alpha\|_L \leq \|\alpha\|_1/2$.

By the reverse triangle inequality we have $\big||\boldsymbol{w}^\top \boldsymbol{x} - b| - |\boldsymbol{w}^\top \boldsymbol{y} - b|\big| \leq \big|\boldsymbol{w}^\top(\boldsymbol{x} - \boldsymbol{y})\big|$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $\boldsymbol{w} \in \mathbb{S}^{d-1}$, $b \in \mathbb{R}$. Therefore, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ we see that

$$|h_\alpha(\boldsymbol{x}) - h_\alpha(\boldsymbol{y})| = \frac{1}{2} \left| \int_{\mathbb{S}^{d-1} \times \mathbb{R}} \big(|\boldsymbol{w}^\top \boldsymbol{x} - b| - |\boldsymbol{w}^\top \boldsymbol{y} - b|\big) \, d\alpha(\boldsymbol{w}, b) \right| \tag{51}$$

$$\leq \frac{1}{2} \int_{\mathbb{S}^{d-1} \times \mathbb{R}} \big||\boldsymbol{w}^\top \boldsymbol{x} - b| - |\boldsymbol{w}^\top \boldsymbol{y} - b|\big| \, d|\alpha|(\boldsymbol{w}, b) \tag{52}$$

$$\leq \frac{1}{2} \int_{\mathbb{S}^{d-1} \times \mathbb{R}} |\boldsymbol{w}^\top(\boldsymbol{x} - \boldsymbol{y})| d|\alpha|(\boldsymbol{w}, b) \tag{53}$$

$$\leq \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{y}\| \|\alpha\|_1 \tag{54}$$

which shows h_α is globally Lipschitz with $\|h_\alpha\|_L \leq \|\alpha\|_1/2$.

More generally, for any infinite-width net $f = h_{\alpha, \boldsymbol{v}, c}$ with $\alpha \in M(\mathbb{S}^{d-1} \times \mathbb{R})$, $\boldsymbol{v} \in \mathbb{R}^d$ and $c \in \mathbb{R}$. By an even/odd decomposition $\alpha = \alpha_+ + \alpha_-$ we have $f = h_{\alpha_+, \boldsymbol{v}_0 + \boldsymbol{v}, c}$, where $\boldsymbol{v}_0 = \frac{1}{2} \int_{\mathbb{S}^{d-1} \times \mathbb{R}} \boldsymbol{w} d\alpha_-(\boldsymbol{w}, b)$. Hence, $\|\boldsymbol{v}_0\|_2 \leq \|\alpha_-\|_1/2$, Therefore, by the triangle inequality, $\|f\|_L \leq \|\alpha_+\|_1/2 + \|\alpha_-\|_1/2 + \|\boldsymbol{v}\| \leq \|\alpha\|_1 + \|\boldsymbol{v}\|$, which gives the claim. □

The following lemma shows every infinite-width net is the limit of a sequence of finite width nets converging uniformly on compact subsets, and will be used in our proof of Theorem 1.

Lemma 4. *Let $f = h_{\alpha, \boldsymbol{v}, c}$ for any $\alpha \in M(\mathbb{P}^d)$,$\boldsymbol{v} \in \mathbb{R}^d$, and $c \in \mathbb{R}$. Then there exists a sequence of discrete measures $\alpha_n \in M(\mathbb{P}^d)$ with $\|\alpha_n\|_1 \leq \|\alpha\|_1$ such that $f_n = h_{\alpha_n, \boldsymbol{v}, c}$ converges uniformly to f on compact subsets of $\mathbb{R}^d$.*

Proof. For any $\alpha \in M(\mathbb{P}^d)$ there exists a sequence of discrete measures $\{\alpha_n\}$ converging *narrowly* to α (meaning $\int \varphi d\alpha_n \to \int \varphi d\alpha$ for all $\varphi \in C_b(\mathbb{S}^{d-1} \times \mathbb{R})$) such that $\|\alpha_n\|_1 \leq \|\alpha\|_1$ (Malliavin, 2012, Chapter 2, Theorem 6.9). Let $f_n = h_{\alpha_n,\boldsymbol{v},c}$. Since the function $(\boldsymbol{w}, b) \mapsto |\boldsymbol{w}^\top \boldsymbol{x} - b| - |b|$ is continuous and bounded, we have $f_n(\boldsymbol{x}) \to f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$, *i.e.*, $f_n \to f$ pointwise.

To show uniform convergence on compact sets, observe that the minimum Lipschitz constant of f_n, denoted by $\|f_n\|_L$, is bounded as $\|f_n\|_L \le \|\alpha_n\|_1 \le \|\alpha\|_1$. This shows the sequence f_n is uniformly Lipschitz. Therefore, by a corollary of the Arzela-Ascoli theorem, the sequence f_n is uniformly convergent on any compact set of $\mathbb{R}^d$, as claimed. □

Finally, we give a lemma deriving equivalent expressions for $\overline{R}(f)$ and $\overline{R}_1(f)$ in terms the limiting norm of sequences of finite-width ReLU networks converging pointwise to f; this formulation will be easier to work with in our proofs.

Lemma 5. *For any f we have,*

$$\overline{R}(f) = \inf\left\{\liminf_{n\to\infty} C(\theta_n) : \theta_n \in \Theta',\ \ g_{\theta_n} \to f \textit{ pointwise},\ \ g_{\theta_n}(\mathbf{0}) = f(\mathbf{0})\ \ \forall n\right\}. \tag{55}$$

and

$$\overline{R}_1(f) = \inf\left\{\liminf_{n\to\infty} C(\theta_n) : \theta_n \in \Theta', \boldsymbol{v}_n \in \mathbb{R}^d,\ \ g_{\theta_n,\boldsymbol{v}_n} \to f \textit{ pointwise},\ \ g_{\theta_n,\boldsymbol{v}_n}(\mathbf{0}) = f(\mathbf{0})\ \ \forall n\right\}. \tag{56}$$

with $\overline{R}(f)$ as defined in (7) *and $\overline{R}_1(f)$ as defined in* (12).

Proof. We prove the identity in (55) for $\overline{R}(f)$; the identity in (56) for $\overline{R}_1(f)$ follows by the same argument.

Let $\widetilde{R}(f)$ denote the right-hand side of (55), and define

$$R_\varepsilon(f) := \inf_{\theta\in\Theta'} C(\theta)\ \ s.t.\ \ |g_\theta(\boldsymbol{x}) - f(\boldsymbol{x})| \le \varepsilon\ \ \forall\ \|\boldsymbol{x}\| \le 1/\varepsilon \text{ and } g_\theta(\mathbf{0}) = f(\mathbf{0}) \tag{57}$$

so that $\overline{R}(f) = \lim_{\varepsilon\to 0} R_\varepsilon(f)$.

First, suppose $\overline{R}(f)$ is finite. Let $\varepsilon_n = 1/n$. Then by definition of $\overline{R}(f)$, for all n there exists $\theta_n \in \Theta'$ such that $f_n := g_{\theta_n}$ satisfies $\|f_n(\boldsymbol{x}) - f(\boldsymbol{x})\| \le \varepsilon_n$, $f_n(\mathbf{0}) = f(\mathbf{0})$, and $C(\theta_n) \le R_{\varepsilon_n}(f_n) + \varepsilon_n$. Hence, $f_n \to f$ pointwise with $f_n(\mathbf{0}) = f(\mathbf{0})$, while

$$\liminf_{n\to\infty} C(\theta_n) \le \liminf_{n\to\infty} (R_{\varepsilon_n}(f_n) + \varepsilon_n) = \overline{R}(f). \tag{58}$$

Therefore, we have shown $\widetilde{R}(f) \le \overline{R}(f)$.

Conversely, suppose $\widetilde{R}(f)$ is finite. Fix any $\varepsilon > 0$. Then by definition of $\widetilde{R}(f)$ there exists a sequence $\theta_n \in \Theta'$ such that $f_n := g_{\theta_n}$ converges to f pointwise, and where $\lim_{n\to\infty} C(\theta_n)$ exists with $\lim_{n\to\infty} C(\theta_n) < \widetilde{R}(f) + \varepsilon$. Note that $f_n = h_{\alpha_n,c_n}$ where $\alpha_n \in M(\mathbb{S}^{d-1} \times \mathbb{R})$ is the discrete measure defined by $\alpha_n(\boldsymbol{w}, b) = \sum_{i=1}^k a_i \delta_{(\boldsymbol{w}_i,b_i)}(\boldsymbol{w}, b)$ for some $c_n \in \mathbb{R}$, and we also have $C(\theta_n) = \|\alpha_n\|_1$. Therefore $\lim_{n\to\infty} \|\alpha_n\|_1 \le \overline{R}(f) + \varepsilon$, and so there exists an N_1 such that for all $n \ge N_1$ we have $\|\alpha_n\|_1 < \overline{R}(f) + \varepsilon$. By Proposition 8, the Lipschitz constant of f_n is bounded above by $\|\alpha_n\|_1$, hence the sequence f_n is uniformly Lipschitz. This implies $f_n \to f$ uniformly on compact subsets, and so there exists an N_2 such that $\|f_n(\boldsymbol{x}) - f(\boldsymbol{x})\| \le \varepsilon$ for all $\|\boldsymbol{x}\| \le 1/\varepsilon$ and $f_n(\mathbf{0}) = f(\mathbf{0})$ for all $n \ge N_2$. Hence for all $n \ge N_2$, f_n belongs to the feasible set in the definition of $R_\varepsilon(\cdot)$. Therefore, for all $n \ge \max\{N_1, N_2\}$ have

$$R_\varepsilon(f) \le C(\theta_n) = \|\alpha_n\|_1 \le \overline{R}(f) + \varepsilon. \tag{59}$$

Finally, taking the limit as $\varepsilon \to 0$, we have $\overline{R}(f) \le \widetilde{R}(f)$. Hence, we have shown $\overline{R}(f)$ is finite if and only if $\widetilde{R}(f)$ is finite, in which case $\overline{R}(f) = \widetilde{R}(f)$. □

C EXTENSION OF $\mathcal{R}$-NORM TO LIPSCHITZ FUNCTIONS AND PROOF OF THEOREM 1

To simplify notation we let $\mathcal{S}(\mathbb{P}^d)$ denote the space of *even* Schwartz functions on $\mathbb{S}^{d-1}\times\mathbb{R}$, *i.e.*, $\psi \in \mathcal{S}(\mathbb{P}^d)$ if $\psi \in C^\infty(\mathbb{S}^{d-1} \times \mathbb{R})$, $\psi(\boldsymbol{w}, b) = \psi(-\boldsymbol{w}, -b)$ for all $(\boldsymbol{w}, b) \in \mathbb{S}^{d-1} \times \mathbb{R}$, and $\psi(\boldsymbol{w}, b)$ and all of its partial derivatives decrease rapidly as $|b| \to \infty$.

We will need the following result characterizing the image of Schwartz functions under the dual Radon transform due to Solmon (1987):

Lemma 6 (Solmon (1987), Theorem 7.7)**.** *Let $\psi \in \mathcal{S}(\mathbb{P}^d)$ and define $\varphi = \gamma_d(-\Delta)^{(d-1)/2}\mathcal{R}^*\{\psi\}$. Then $\varphi \in C^\infty(\mathbb{R}^d)$ with $\varphi(\boldsymbol{x}) = O(\|\boldsymbol{x}\|^{-d})$ and $\Delta\varphi(\boldsymbol{x}) = O(\|\boldsymbol{x}\|^{-d-2})$ as $\|\boldsymbol{x}\| \to \infty$. Moreover, $\mathcal{R}\{\varphi\} = \psi$.*

Using the above result we show the functional $\|f\|_{\mathcal{R}}$ given in Definition 1 is well-defined:

Proposition 9. *For any $f \in \mathrm{Lip}(\mathbb{R}^d)$, the linear functional $L_f(\psi) := \gamma_d\langle f, (-\Delta)^{(d+1)/2}\mathcal{R}^*\{\psi\}\rangle$ is finite for all $\psi \in \mathcal{S}(\mathbb{P}^d)$, hence*

$$\|f\|_{\mathcal{R}} = \sup\left\{L_f(\psi) : \psi \in \mathcal{S}(\mathbb{P}^d), \|\psi\|_\infty \leq 1\right\} \tag{60}$$

is a well-defined functional taking on values in $[0, +\infty]$.

Proof. Since f is globally Lipschitz we have $|f(\boldsymbol{x})| = O(\|\boldsymbol{x}\|)$, while for any $\psi \in \mathcal{S}(\mathbb{P}^d)$ by Lemma 6 we have $|(-\Delta)^{(d+1)/2}\mathcal{R}^*\{\psi\}| = O(\|\boldsymbol{x}\|^{-d-2})$, hence $|f(\boldsymbol{x})(-\Delta)^{(d+1)/2}\mathcal{R}^*\{\psi\}(\boldsymbol{x})| = O(\|\boldsymbol{x}\|^{-d-1})$ is absolutely integrable, and so $\langle f, (-\Delta)^{(d+1)/2}\mathcal{R}^*\{\psi\}\rangle$ is finite. If $\langle f, (-\Delta)^{(d+1)/2}\mathcal{R}^*\{\psi\}\rangle \neq 0$, we can choose the sign of ψ so that the inner product is positive, which shows that $\|f\|_{\mathcal{R}} \geq 0$. □

In Section 4 we showed $\Delta h_\alpha = \mathcal{R}^*\{\alpha\}$ when α was a measure with a smooth density having rapid decay. The next lemma, which is key to proving Theorem 1, shows this equality still holds in the sense of distributions when α is any measure in $M(\mathbb{P}^d)$.

Lemma 7. *Let $f = h_{\alpha,\boldsymbol{v},c}$ for any $\alpha \in M(\mathbb{P}^d), \boldsymbol{v} \in \mathbb{R}^d, c \in \mathbb{R}$. Then we have $\langle f, \Delta\varphi\rangle = \langle \alpha, \mathcal{R}\{\varphi\}\rangle$ for all $\varphi \in C^\infty(\mathbb{R}^d)$ such that $\varphi(\boldsymbol{x}) = O(\|\boldsymbol{x}\|^{-d})$ and $\Delta\varphi(\boldsymbol{x}) = O(\|\boldsymbol{x}\|^{-d-2})$ as $\|\boldsymbol{x}\| \to \infty$.*

Proof. Consider the ridge function $r_{\boldsymbol{w},b}(\boldsymbol{x}) := \frac{1}{2}|\boldsymbol{w}^\top\boldsymbol{x} - b|$, which is generated by the even measure $\alpha_0(\boldsymbol{w}', b') = \frac{1}{2}(\delta(\boldsymbol{w}' - \boldsymbol{w}, b' - b) + \delta(\boldsymbol{w}' + \boldsymbol{w}, b' + b))$. An easy calculation shows that $\Delta r_{\boldsymbol{w},b}(x) = \delta(\boldsymbol{w}^\top\boldsymbol{x} - b)$ in the sense of distributions, *i.e.*,

$$\int r_{\boldsymbol{w},b}(\boldsymbol{x})\Delta\varphi(\boldsymbol{x})\,d\boldsymbol{x} = \int_{\boldsymbol{w}^\top\boldsymbol{x}=b} \varphi(\boldsymbol{x})\,d\boldsymbol{x} = \mathcal{R}\{\varphi\}(\boldsymbol{w}, b) \tag{61}$$

for all test functions $\varphi \in \mathcal{S}(\mathbb{R}^d)$. Since $\mathcal{R}\{\varphi\}(\boldsymbol{w}, b)$ is finite valued for all $\varphi \in C^\infty(\mathbb{R}^d)$ with decay like $O(\|\boldsymbol{x}\|^{-d})$, by continuity $\Delta r_{\boldsymbol{w},b}(\boldsymbol{x})$ extends uniquely to a distribution acting on this larger space of test functions.

Now consider the more general case of $f = h_\alpha$ with $\alpha \in M(\mathbb{P}^d)$. Then for all $\varphi \in C^\infty(\mathbb{R}^d)$ with $|\varphi(x)| = O(\|x\|^{-d})$ and $|\Delta\varphi(x)| = O(\|x\|^{-d-2})$ as $\|x\| \to \infty$ we have

$$\begin{aligned}
\int_{\mathbb{R}^d} f(x)\Delta\varphi(x)\,dx &= \int_{\mathbb{R}^d}\left(\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \frac{1}{2}(|w^\top x - b| - |b|)\,d\alpha(w, b)\right)\Delta\varphi(x)\,dx && (62)\\
&= \int_{\mathbb{S}^{d-1}\times\mathbb{R}}\left(\int_{\mathbb{R}^d} \frac{1}{2}(|w^\top x - b| - |b|)\Delta\varphi(x)\,dx\right)d\alpha(w, b) && (63)\\
&= \int_{\mathbb{S}^{d-1}\times\mathbb{R}}\left(\int_{\mathbb{R}^d} r_{w,b}(x)\Delta\varphi(x)\,dx\right)d\alpha(w, b) && (64)\\
&= \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \mathcal{R}\{\varphi\}(w, b)\,d\alpha(w, b) && (65)
\end{aligned}$$

where in (63) we applied Fubini's theorem to exchange the order of integration, whose application is justified since

$$h_+(x) := \frac{1}{2}\int_{\mathbb{S}^{d-1}\times\mathbb{R}} (|w^\top x - b| - |b|)\,d|\alpha|(w, b) \leq \|\alpha\|_1\|x\| \tag{66}$$

and by assumption $|\Delta\varphi(x)| = O(\|x\|^{-d-2})$, hence $h_+(x)|\Delta\varphi(x)| = O(\|x\|)^{-d-1}$, and so $\int h_+(x)|\Delta\varphi(x)|\,dx < \infty$.

Finally, if $f = h_{\alpha,\boldsymbol{v},c}$ for any $\alpha \in M(\mathbb{P}^d)$, $\boldsymbol{v} \in \mathbb{R}^d$, $c \in \mathbb{R}$, since affine functions vanish under the Laplacian we have $\langle f, \Delta\varphi\rangle = \langle h_\alpha, \Delta\varphi\rangle$, reducing this to the previous case, which gives the claim. □

The following lemma shows the $\|f\|_{\mathcal{R}}$ is finite if and only if f is an infinite width net, in which case $\|f\|_{\mathcal{R}}$ is given by the total variation norm of the unique even measure defining f.

Lemma 8. *Let $f \in \mathrm{Lip}(\mathbb{R}^d)$, and define the linear functional $L_f(\psi) = \gamma_d\langle f, (-\Delta)^{(d-1)/2}\mathcal{R}^*\{\psi\}\rangle$ for all $\psi \in \mathcal{S}(\mathbb{P}^d)$. Then $\|f\|_{\mathcal{R}}$ is finite if and only if there exists a unique measure $\alpha \in M(\mathbb{P}^d)$ such that $L_f(\psi) = \int \psi\, d\alpha$ for all $\psi \in \mathcal{S}(\mathbb{P}^d)$ and unique $\boldsymbol{v} \in \mathbb{R}^d, c \in \mathbb{R}$ with $f = h_{\alpha,\boldsymbol{v},c}$, and $\|f\|_{\mathcal{R}} = \|\alpha\|_1$.*

Proof. Suppose $\|f\|_{\mathcal{R}}$ is finite. Then by definition f belongs to $\mathrm{Lip}(\mathbb{R}^d)$ and the linear functional L_f is continuous on $\mathcal{S}(\mathbb{P}^d)$ with norm $\|f\|_{\mathcal{R}}$. Since $\mathcal{S}(\mathbb{P}^d)$ is a dense subspace of $C_0(\mathbb{P}^d)$, by continuity there exists a unique extension $\tilde{L}_f$ to all of $C_0(\mathbb{P}^d)$ with the same norm. Hence, by the Riesz representation theorem, there is a unique measure $\alpha \in M(\mathbb{P}^d)$ such that $\tilde{L}_f(\psi) = \int \psi\, d\alpha$ for all $\psi \in C_0(\mathbb{P}^d)$ and $\|f\|_{\mathcal{R}} = \|\alpha\|_1$.

We now show $f = h_{\alpha,\boldsymbol{v},c}$ for some $\boldsymbol{v} \in \mathbb{R}^d, c \in \mathbb{R}$. First, we prove $\Delta f = \Delta h_\alpha$ as tempered distributions (*i.e.*, as linear functionals on the space of Schwartz functions $\mathcal{S}(\mathbb{R}^d)$). By Lemma 7 we have $\langle \Delta h_\alpha, \varphi\rangle = \langle \alpha, \mathcal{R}\{\varphi\}\rangle$ for any $\varphi \in \mathcal{S}(\mathbb{R}^d)$, hence

$$\begin{aligned}
\langle \Delta h_\alpha, \varphi\rangle &= \langle \alpha, \mathcal{R}\{\varphi\}\rangle && (67)\\
&= \tilde{L}_f(\mathcal{R}\{\varphi\}) && (68)\\
&= L_f(\mathcal{R}\{\varphi\}) && (69)\\
&= \gamma_d\langle f, (-\Delta)^{(d+1)/2}\mathcal{R}^*\{\mathcal{R}\{\varphi\}\}\rangle && (70)\\
&= -\gamma_d\langle f, \Delta(-\Delta)^{(d-1)/2}\mathcal{R}^*\{\mathcal{R}\{\varphi\}\}\rangle && (71)\\
&= \langle f, \Delta\varphi\rangle && (72)\\
&= \langle \Delta f, \varphi\rangle && (73)
\end{aligned}$$

where in (69) we used the fact that $\mathcal{R}\{\varphi\} \in \mathcal{S}(\mathbb{P}^d)$ for all $\varphi \in \mathcal{S}(\mathbb{R}^d)$ (Helgason, 1999, Theorem 2.4), and in (72) we used the inversion formula for Radon transform: $-\gamma_d(-\Delta)^{(d-1)/2}\mathcal{R}^*\{\mathcal{R}\{\varphi\}\} = \varphi$ for all $\varphi \in \mathcal{S}(\mathbb{R}^d)$ (Helgason, 1999, Theorem 3.1).

Hence, we have shown $\Delta f = \Delta h_\alpha$ as tempered distributions. This means $f - h_\alpha$ is in null space of the Laplacian acting on tempered distributions, which implies $f - h_\alpha = p$ where p is some harmonic polynomial (*i.e.*, p is a polynomial in $x = (x_1, ..., x_d)$ such that $\Delta p(x) = 0$ for all $x \in \mathbb{R}^d$). Finally, since both f and h_α are Lipschitz they have at most linear growth at infinity, so must p. This implies p must be an affine function $p(\boldsymbol{x}) = \boldsymbol{v}^\top \boldsymbol{x} + c$, which shows $f = h_{\alpha,\boldsymbol{v},c}$ as claimed.

Conversely, suppose $f = h_{\alpha,\boldsymbol{v},c}$ for some $\alpha \in M(\mathbb{P}^d), \boldsymbol{v} \in \mathbb{R}^d, c \in \mathbb{R}$. Let $\psi \in \mathcal{S}(\mathbb{P}^d)$. By Lemma 6, the function $\varphi = -\gamma_d(-\Delta)^{(d-1)/2}\mathcal{R}^*\{\psi\}$ is in $C^\infty(\mathbb{R}^d)$ with $\varphi(x) = O(\|x\|^{-d})$, $\Delta\varphi(x) = O(\|x\|^{-d-2})$ as $\|x\| \to \infty$, and $\psi = \mathcal{R}\{\varphi\}$. Hence, by Lemma 7 we have

$$L_f(\psi) = \langle f, \Delta\varphi\rangle = \langle \alpha, \mathcal{R}\{\varphi\}\rangle = \langle \alpha, \psi\rangle. \tag{74}$$

This shows

$$\begin{aligned}
\|f\|_{\mathcal{R}} &= \sup\{\langle \alpha, \psi\rangle : \psi \in \mathcal{S}(\mathbb{P}^d), \|\psi\|_\infty \leq 1\} && (75)\\
&= \sup\{\langle \alpha, \psi\rangle : \psi \in C_0(\mathbb{P}^d), \|\psi\|_\infty \leq 1\} && (76)\\
&= \|\alpha\|_1 && (77)
\end{aligned}$$

where the second to last equality holds since $\mathcal{S}(\mathbb{R}^d)$ is a dense subspace of $C_0(\mathbb{R}^d)$, and the last equality is by the dual characterization of the total variation norm. □

Now we give the proof of our main theorem.

Proof of Theorem 1. Suppose $\overline{R}_1(f)$ is finite. Then by Lemma 5 there exists a sequence of finite width ReLU nets $f_n(\boldsymbol{x}) := \sum_{i=1}^{k_n} a_{n,i}[\boldsymbol{w}_{n,i}^\top \boldsymbol{x} - b_{n,i}]_+ + \boldsymbol{v}_n^\top \boldsymbol{x} + c_n$ that converge pointwise to f. Let α_n be the discrete measure $\alpha_n = \sum_{i=1}^{k_n} a_{n,i}\delta_{(\boldsymbol{w}_{n,i},b_{n,i})}$, so that we can write $f_n(\boldsymbol{x}) =$

$\int_{\mathbb{S}^{d-1}\times\mathbb{R}}[\boldsymbol{w}^\top\boldsymbol{x} - b]_+ d\alpha_n(\boldsymbol{w}, b) + \boldsymbol{v}_n^\top\boldsymbol{x} + c_n$. Then $C(\theta_n) = \|\alpha_n\|_1$, and by finiteness of $\overline{R}_1(f)$ there exists an upper bound $B > 0$ such that $\|\alpha_n\|_1 \leq B$ for all n, while $f_n(\mathbf{0}) = f(\mathbf{0})$.

Since $\overline{R}_1(f)$ is invariant to additive constants, without loss of generality assume $f(\mathbf{0}) = 0$. Hence, $f_n(\mathbf{0}) = \int_{\mathbb{S}^{d-1}\times\mathbb{R}}[-b]_+ d\alpha_n(\boldsymbol{w}, b) + c_n = 0$ and so $c_n = \int_{\mathbb{S}^{d-1}\times\mathbb{R}}[-b]_+ d\alpha_n(\boldsymbol{w}, b)$, which gives $f_n(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}\times\mathbb{R}}([\boldsymbol{w}^\top\boldsymbol{x} - b]_+ - [-b]_+) d\alpha_n(\boldsymbol{w}, b) + \boldsymbol{v}_n^\top\boldsymbol{x}$. Finally, letting α_n^+ be the even part of α, we have $f_n = h_{\alpha_n^+, \boldsymbol{v}_n', 0}$ for some $\boldsymbol{v}_n' \in \mathbb{R}^d$.

First we show the pointwise convergence of f_n implies α_n^+ converges *weakly* to a measure $\alpha \in M(\mathbb{P}^d)$ such that $\|f\|_{\mathcal{R}} = \|\alpha\|_1$, meaning $\int \varphi \, d\alpha_n^+ \to \int \varphi \, d\alpha$ for all $\varphi \in C_0(\mathbb{P}^d)$. To prove weak convergence of α_n^+ it suffices to show there exists a dense set $D \subset C_0(\mathbb{P}^d)$ such that $\int \varphi \, d\alpha_n^+$ converges for all $\varphi \in D$ (Malliavin, 2012, Chapter 2, Theorem 6.8); we will do this for $D = \mathcal{S}(\mathbb{P}^d)$ using the range characterization for the dual Radon transform given in Lemma 6.

Fix any $\psi \in \mathcal{S}(\mathbb{P}^d)$ and set $\varphi = \gamma_d(-\Delta)^{(d-1)/2}\mathcal{R}^*\{\psi\}$. Let $g_n = f_n\Delta\varphi$ and $g = f\Delta\varphi$ so that $g_n \to g$ pointwise. Also, since we have the bound $|f_n(\boldsymbol{x})| \leq \|\alpha_n^+\|_1\|\boldsymbol{x}\| \leq B\|\boldsymbol{x}\|$, and by Lemma 6 we know $|\Delta\varphi(\boldsymbol{x})| = O(\|\boldsymbol{x}\|^{-d-2})$ we have $|g_n(\boldsymbol{x})| \leq B\|\boldsymbol{x}\||\Delta\varphi(\boldsymbol{x})| = O(\|\boldsymbol{x}\|^{-d-1})$ is integrable. Therefore by the Lebesgue dominated convergence theorem

$$\lim_{n\to\infty} \int f_n \, \Delta\varphi \, d\boldsymbol{x} = \int f \, \Delta\varphi \, d\boldsymbol{x} < \infty. \tag{78}$$

By Lemma 7, and the fact that $\mathcal{R}\{\varphi\} = \psi$, we have

$$\int f_n \, \Delta\varphi \, d\boldsymbol{x} = \int \mathcal{R}\{\varphi\} \, d\alpha_n^+ = \int \psi \, d\alpha_n^+. \tag{79}$$

Hence,

$$\lim_{n\to\infty} \int \psi d\alpha_n^+ = \int f \, \Delta\varphi \, d\boldsymbol{x}. \tag{80}$$

Since $\mathcal{S}(\mathbb{P}^d)$ is dense in $C_0(\mathbb{P}^d)$, and $\|\alpha_n^+\|_1 \leq \|\alpha_n\|_1 \leq B$, this shows α_n^+ converges weakly to a measure $\alpha \in M(\mathbb{P}^d)$.

Now, by definition of weak convergence we have

$$\int \psi \, d\alpha = \lim_{n\to\infty} \int \psi \, d\alpha_n^+ = \int f \, \Delta\varphi \, d\boldsymbol{x} = -\gamma_d \int f \, (-\Delta)^{(d+1)/2}\mathcal{R}^*\{\psi\} \, d\boldsymbol{x} \tag{81}$$

This shows the functional $L_f(\psi) = \gamma_d\langle f, (-\Delta)^{(d+1)/2}\mathcal{R}^*\{\psi\}\rangle = \int \psi d\alpha$ for all $\psi \in \mathcal{S}(\mathbb{P}^d)$. Therefore, by uniqueness of α in Lemma 8, we have $\|f\|_{\mathcal{R}} = \|\alpha\|_1$ and $f = h_{\alpha,\boldsymbol{v},0}$ for some $\boldsymbol{v} \in \mathbb{R}^d$.

Finally, note that $\alpha_n^+ \to \alpha$ weakly implies $\|\alpha\|_1 \leq \liminf_{n\to\infty}\|\alpha_n^+\|_1 \leq \liminf_{n\to\infty}\|\alpha_n\|_1$. Since α_n was an arbitrary sequence from the infimum over all such sequences, we have $\overline{R}_1(f) \geq \|\alpha\|_1 = \|f\|_{\mathcal{R}}$. Also, by Lemma 4 we know that there exists a sequence of discrete $\alpha_n' \in M(\mathbb{P}^d)$ with $\|\alpha_n'\|_1 \leq \|\alpha\|_1$ such that the sequence of finite width nets $f_n' = h_{\alpha_n',\boldsymbol{v},0}$ converges pointwise to f with $f_n'(\mathbf{0}) = f(\mathbf{0})$, and so $\overline{R}_1(f) \leq \|\alpha\|_1$. Hence we have shown $\overline{R}_1(f) = \|\alpha\|_1 = \|f\|_{\mathcal{R}}$. □

D Optimization characterization of $\overline{R}(f)$ and $\overline{R}_1(f)$

Here we establish the optimization equivalents of $\overline{R}(f)$ and $\overline{R}_1(f)$ given in (9) and (13).

First, we prove a lemma:

Lemma 9. *Suppose $f = h_{\alpha,\boldsymbol{v},c}$ for some $\alpha \in M(\mathbb{S}^{d-1}\times\mathbb{R})$, $\boldsymbol{v} \in \mathbb{R}^d$, $c \in \mathbb{R}$. If $f = h_{\beta,\boldsymbol{v}',c'}$ for any other $\beta \in M(\mathbb{S}^{d-1}\times\mathbb{R})$, $\boldsymbol{v}' \in \mathbb{R}^d$, $c' \in \mathbb{R}$ then $\alpha^+ = \beta^+$ where α^+ and β^+ are the even parts of α and β, respectively.*

Proof. Note that $f = h_{\alpha,c} = h_{\alpha_+} + h_{\alpha_-,\boldsymbol{v},c} = h_{\alpha_+,\boldsymbol{v}_1,c}$ for some $\boldsymbol{v}_1 \in \mathbb{R}^d$. Similarly, $f = h_{\beta,\boldsymbol{v}',d} = h_{\beta_+,\boldsymbol{v}_1',d}$ for some $\boldsymbol{v}_1' \in \mathbb{R}^d$. Therefore, by the uniqueness of the even mesaure generating an infinite-width net established in Lemma 8, we have $\alpha_+ = \beta_+$ □

We now prove the equivalence of $\overline{R}(f)$ with (9). (Note: the following proof reuses many pieces from the proof of Theorem 1 given above in Appendix C.)

Proof. By Lemma 9, the minimization problem (9) is equivalent to

$$M(f) := \min_{\alpha_- \text{ odd}, c} \|\alpha_+ + \alpha_-\|_1 \;\; s.t. \;\; f = h_{\alpha_+} + h_{\alpha_-, c}, \tag{82}$$

First we show $M(f) \leq \overline{R}(f)$. Assume $\overline{R}(f)$ is finite. Without loss of generality we assume $f(\mathbf{0}) = 0$. Following the same steps as in the proof of Theorem 1, we can show there exists a sequence of discrete measures $\alpha_n \in M(\mathbb{S}^{d-1} \times \mathbb{R})$ such that $f_n = h_{\alpha_n}$ with $f_n \to f$ pointwise, and such that $\|\alpha_n\|_1 \leq B$. Letting $\alpha_n = \alpha_n^+ + \alpha_n^-$ be the even/odd decomposition, we have $f_n = h_{\alpha_n^+, \boldsymbol{v}_n}$ where $\boldsymbol{v}_n = \frac{1}{2}\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \boldsymbol{w} \, d\alpha_n^-(\boldsymbol{w}, b)$.

By the same steps in the proof of Theorem 1, we can show this implies α_n^+ converges weakly to a unique even measure α^+ and $f = h_{\alpha^+, \boldsymbol{v}, 0}$ for some $\boldsymbol{v} \in \mathbb{R}^d$. Let α^- be any odd measure such that $\boldsymbol{v} = \frac{1}{2}\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \boldsymbol{w} \, d\alpha^-(\boldsymbol{w}, b)$ and set $\alpha = \alpha^+ + \alpha^-$. Then we have $f = h_{\alpha, 0}$. By similar argument to that in Theorem 1, this implies $M(f) \leq \overline{R}(f)$.

Finally, by Lemma 4 there exists a sequence of discrete measures $\alpha_n' \in M(\mathbb{S}^{d-1} \times \mathbb{R})$ (corresponding to parameters $\theta_n' \in \Theta'$) converging narrowly to α, while $h_{\alpha_n'}$ converges to f pointwise, with $C(\theta_n') = \|\alpha_n'\|_1 \leq \|\alpha\|_1$. Since the choice of α was arbitrary, we see that $\overline{R}(f) \leq M(f)$.

□

Now we show that $\overline{R}_1(f)$ is also equivalent to the optimization problem in (13).

Proposition 10. *Suppose $f = h_{\alpha, \boldsymbol{v}, c}$ for some $\alpha \in M(\mathbb{S}^{d-1} \times \mathbb{R})$, $\boldsymbol{v} \in \mathbb{R}^d$, $c \in \mathbb{R}$. Write $\alpha = \alpha_+ + \alpha_-$ where α_+ is even and α_- is odd. Then*

$$\|\alpha_+\|_1 = \min_{\beta, \boldsymbol{v}', c'} \|\beta\|_1 \;\; s.t. \;\; f = h_{\beta, \boldsymbol{v}', c'} \tag{83}$$

and so $\overline{R}_1(f) = \|\alpha_+\|_1 = \|f\|_{\mathcal{R}}$.

Proof. By Lemma 9, the minimization problem (10) reduces to

$$\min_{\alpha_- \text{ odd}, \boldsymbol{v}', c'} \|\alpha_+ + \alpha_-\|_1 \;\; s.t. \;\; f = h_{\alpha_+} + h_{\alpha_-, \boldsymbol{v}', c'}, \tag{84}$$

By Proposition 7, we have $\|\alpha_+\|_1 \leq \|\alpha_+ + \alpha_-\|_1$ for any α_- odd. Since $f = h_{\alpha+, \boldsymbol{v}_1, c}$, we see that a global minimum is reached at $\alpha_- = 0$, $\boldsymbol{v}' = \boldsymbol{v}_1$, and $c' = c$, which gives the objective value $\|\alpha_+\|_1 = \|f\|_{\mathcal{R}}$. □

Proof of Lemma 1 Let $f(\boldsymbol{x}) = \sum_{i=1}^k a_i [\boldsymbol{w}_i^\top \boldsymbol{x} - b_i]_+ + \boldsymbol{v}^\top \boldsymbol{x} + c$. Then $f = h_{\alpha^+, \boldsymbol{v}', c'}$ for some $\boldsymbol{v}' \in \mathbb{R}^d$ and $c' \in \mathbb{R}$ where α^+ is the even measure

$$\alpha^+(\boldsymbol{w}, b) = \sum_{i=1}^k \frac{a_i}{2} \left(\delta(\boldsymbol{w} - \boldsymbol{w}_i, b - b_i) + \delta(\boldsymbol{w} + \boldsymbol{w}_i, b + b_i)\right) \tag{85}$$

Hence, by Proposition 10, and using the fact that the $(\boldsymbol{w}_i, b_i)$ are distinct under (under the identification $(\boldsymbol{w}, b) = (-\boldsymbol{w}, -b)$), we have

$$\overline{R}_1(f) = \|f\|_{\mathcal{R}} = \|\alpha^+\|_1 = \sum_{i=1}^k |a_i| \tag{86}$$

E Proof of Theorem 2

We show how our results change without the addition of the unregularized linear unit $\boldsymbol{v}^\top \boldsymbol{x}$ in (3). Specifically, we want to characterize $\overline{R}(f)$ given in (7) (or equivalently its optimization formulation (9)). Unlike in the 1D setting, $\overline{R}(f)$ does not have a simple closed form expression in higher dimensions. However, for any $f \in \mathrm{Lip}(\mathbb{R}^d)$ we prove the bounds

$$\max\{\|f\|_{\mathcal{R}}, 2\|\nabla f(\infty)\|_2\} \leq \overline{R}(f) \leq \|f\|_{\mathcal{R}} + 2\|\nabla f(\infty)\|_2 \tag{87}$$

where $\nabla f(\infty) \in \mathbb{R}^d$ can be thought of as the gradient of the function f "at infinity"; see below for a formal definition. In particular, if $f(\boldsymbol{x})$ vanishes at infinity then $\nabla f(\infty) = 0$ and we have $\overline{R}(f) = \|f\|_{\mathcal{R}} = \overline{R}_1(f)$.

For any $f \in \mathrm{Lip}(\mathbb{R}^d)$, define $\nabla f(\infty) \in \mathbb{R}^d$ by

$$\nabla f(\infty) := \lim_{r\to\infty} \frac{1}{c_d r^{d-1}} \oint_{\|\boldsymbol{x}\|=r} \nabla f(\boldsymbol{x})\, ds(\boldsymbol{x}), \tag{88}$$

We relate $\nabla f(\infty)$ to the "linear part" of an infinite width net. Towards this end, define $\mathcal{V} : M(\mathbb{S}^{d-1} \times \mathbb{R}) \to \mathbb{R}^d$ to be the linear operator given by

$$\mathcal{V}(\alpha) = \frac{1}{2} \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \boldsymbol{w}\, d\alpha(\boldsymbol{w}, b). \tag{89}$$

Note that if $\alpha = \alpha_+ + \alpha_-$ where α_+ is even and α_- is odd, then $\mathcal{V}(\alpha) = \mathcal{V}(\alpha_-)$ since $\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \boldsymbol{w}\, d\alpha_+(\boldsymbol{w}, b) = 0$. In particular, if we set $\boldsymbol{v}_0 = \mathcal{V}(\alpha_-)$, then $h_{\alpha_-}(\boldsymbol{x}) = \boldsymbol{v}_0^\top \boldsymbol{x}$.

Lemma 10. *Suppose $f = h_{\alpha,c}$ for any $\alpha \in M(\mathbb{S}^{d-1} \times \mathbb{R})$, $c \in \mathbb{R}$. Then, $\nabla f(\infty) = \mathcal{V}(\alpha)$.*

Proof. A simple calculation shows the weak gradient of $f = h_{\alpha,c}$ is given by

$$\nabla f(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}\times\mathbb{R}} H(\boldsymbol{w}^\top \boldsymbol{x} - b)\boldsymbol{w}\, d\alpha(\boldsymbol{w}, b) \tag{90}$$

where H is defined as $H(t) = 1$ if $t \geq 0$ and $H(t) = 0$ if $t < 0$ otherwise. Therefore, we have

$$\begin{aligned}
\lim_{r\to\infty} \frac{1}{r^{d-1}} \oint_{\|\boldsymbol{x}\|=r} \nabla f(\boldsymbol{x})\, ds(\boldsymbol{x}) &= \lim_{r\to\infty} \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \int_{\mathbb{S}^{d-1}} H(r\boldsymbol{w}^\top \boldsymbol{w}' - b)\boldsymbol{w}\, d\boldsymbol{w}' d\alpha(\boldsymbol{w}, b) && (91)\\
&= \lim_{r\to\infty} \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \boldsymbol{w} \left(\int_{\boldsymbol{w}^\top \boldsymbol{w}' \geq b/r} d\boldsymbol{w}' \right) d\alpha(\boldsymbol{w}, b) && (92)\\
&= \left(\frac{1}{2} \int_{\mathbb{S}^{d-1}} d\boldsymbol{w}' \right) \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \boldsymbol{w}\ d\alpha(\boldsymbol{w}, b) && (93)
\end{aligned}$$

Finally, dividing both sides by $c_d = \int_{\mathbb{S}^{d-1}} d\boldsymbol{w}$ gives the result. □

Lemma 11. *If $f(\boldsymbol{x}) = \boldsymbol{v}_0^\top \boldsymbol{x} + c$ then $\overline{R}(f) = 2\|\boldsymbol{v}_0\|$.*

Proof. Note that $f = h_{\alpha,c}$ only if α is odd and $\mathcal{V}(\alpha) = \boldsymbol{v}_0$. Hence, we have

$$\overline{R}(f) = \min_{\alpha \text{ odd}} \|\alpha\|_1 \ \ s.t.\ \ \mathcal{V}(\alpha) = \boldsymbol{v}_0 \tag{94}$$

The adjoint $\mathcal{V}^* : \mathbb{R}^d \to C_b(\mathbb{S}^{d-1} \times \mathbb{R})$ is given by $[\mathcal{V}^*\boldsymbol{y}](\boldsymbol{w}, b) = \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{y}$. The dual of the convex program above is given by

$$\max_{\substack{\boldsymbol{y}\in\mathbb{R}^d \\ \|\mathcal{V}^*\boldsymbol{y}\|_\infty \leq 1}} \boldsymbol{v}_0^\top \boldsymbol{y} = \max_{\|\boldsymbol{y}\|\leq 2} \boldsymbol{v}_0^\top \boldsymbol{y} = 2\|\boldsymbol{v}_0\| \tag{95}$$

where we used the fact that $\|\mathcal{V}^*\boldsymbol{y}\|_\infty = \max_{\boldsymbol{w}\in\mathbb{S}^{d-1}} \frac{1}{2}\|\boldsymbol{w}^\top \boldsymbol{y}\| \leq 1$ holds if and only if $\|\boldsymbol{y}\| \leq 2$. This means $2\|\boldsymbol{v}_0\|$ is a lower bound for $\overline{R}(f)$. Since this bound is reached with the primal feasible choice α defined by

$$\alpha(\boldsymbol{w}, b) = \|\boldsymbol{v}_0\| \left(\delta\left(\boldsymbol{w} - \frac{\boldsymbol{v}_0}{\|\boldsymbol{v}_0\|}, b \right) - \delta\left(\boldsymbol{w} + \frac{\boldsymbol{v}_0}{\|\boldsymbol{v}_0\|}, b \right) \right) \tag{96}$$

we have $\overline{R}(f) = 2\|\boldsymbol{v}_0\|$ as claimed. □

Now we give the proof of Theorem 2.

Proof of Theorem 2. Suppose $\|f\|_{\mathcal{R}}$ is finite. Then by Lemma 8, there is a unique even measure α_+ such that $f = h_{\alpha_+,\boldsymbol{v}_0,c}$ for some unique $\boldsymbol{v}_0 \in \mathbb{R}^d, c \in \mathbb{R}$, with $\|f\|_{\mathcal{R}} = \|\alpha_+\|_1$. Therefore, $\overline{R}(f)$ is equivalent to the optimization problem

$$\overline{R}(f) = \min_{\alpha_- \text{ odd}} \|\alpha_+ + \alpha_-\|_1 \;\; s.t. \;\; \mathcal{V}(\alpha_-) = \boldsymbol{v}_0 \tag{97}$$

Since $\|\alpha_+ + \alpha_-\|_1 \leq \|\alpha_+\|_1 + \|\alpha_-\|_1$, by Lemma 11 we see that $\overline{R}(f) \leq \|\alpha_+\|_1 + 2\|\boldsymbol{v}_0\|_2$. Now we show the lower bound. The above optimization problem is equivalent to

$$\overline{R}(f) = \min_{\alpha} \|\alpha\|_1 \;\; s.t. \;\; \mathcal{V}(\alpha) = \boldsymbol{v}_0, \;\; \mathcal{E}(\alpha) = \alpha_+ \tag{98}$$

where $\mathcal{E}(\alpha)$ projects onto the even part of α. Note the Banach space adjoint of $\mathcal{E}^* : C_b(\mathbb{S}^{d-1} \times \mathbb{R}) \to C_b(\mathbb{S}^{d-1} \times \mathbb{R})$ is also projection onto the even part, *i.e.*, $[\mathcal{E}^*\varphi](\boldsymbol{w}, b) = \frac{1}{2}(\varphi(\boldsymbol{w}, b) + \varphi(-\boldsymbol{w}, -b))$. Therefore, the dual problem is given by

$$\sup_{\substack{\varphi \in C_b(\mathbb{S}^{d-1}\times\mathbb{R}), \boldsymbol{y} \in \mathbb{R}^d \\ \|\mathcal{W}^*\boldsymbol{y} + \mathcal{E}^*\varphi\|_\infty \leq 1}} \boldsymbol{v}_0^\top \boldsymbol{y} + \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \varphi(\boldsymbol{w}, b) d\alpha_+(\boldsymbol{w}, b) \tag{99}$$

We can constrain φ to be even without changing the maximum since α_+ is even. Thus the dual feasible set reduces to pairs $(\varphi, \boldsymbol{y})$ with $\varphi \in C_b(\mathbb{S}^{d-1} \times \mathbb{R})$ even and $\boldsymbol{y} \in \mathbb{R}^d$ are such that $|\varphi(\boldsymbol{w}, b) + \frac{1}{2}\boldsymbol{w}^\top y| \leq 1$ for all $(\boldsymbol{w}, b)$.

Taking the supremum over all dual feasible pairs $(\varphi, \boldsymbol{0})$ such that $\|\varphi\|_\infty \leq 1$, we see $\overline{R}(f) \geq \|\alpha_+\|_1 = \|f\|_{\mathcal{R}}$. Likewise, if we choose the dual feasible pair $(\varphi, \boldsymbol{y}) = (0, 2\boldsymbol{v}_0/\|\boldsymbol{v}_0\|)$ then the dual objective is $2\|\boldsymbol{v}_0\|$, hence $\overline{R}(f) \geq 2\|\boldsymbol{v}_0\|$. This gives $\overline{R}(f) \geq \max\{\|f\|_{\mathcal{R}}, 2\|\boldsymbol{v}_0\|\}$, as desired. □

Finally, we show there are examples where the upper bound in Theorem 2 is attained.

Proposition 11. *There exist infinite nets $f : \mathbb{R}^d \to \mathbb{R}$ in all dimensions d such that*

$$\overline{R}(f) = \|f\|_{\mathcal{R}} + 2\|\nabla f(\infty)\|, \tag{100}$$

For example, this holds for $f(x, y) = x + |y|$ in $d = 2$.

Proof. Let $\boldsymbol{w}_+, \boldsymbol{w}_- \in \mathbb{S}^{d-1}$ be orthogonal. Consider $f = h_{\alpha,0}$ defined by $\alpha = \alpha_+ + \alpha_-$ with

$$\alpha_+ = \delta(\boldsymbol{w} - \boldsymbol{w}_+, b) + \delta(\boldsymbol{w} + \boldsymbol{w}_+, b) \tag{101}$$

$$\alpha_- = \delta(\boldsymbol{w} - \boldsymbol{w}_-, b) - \delta(\boldsymbol{w} + \boldsymbol{w}_-, b) \tag{102}$$

Hence, $f(\boldsymbol{x}) = |\boldsymbol{w}_+^\top \boldsymbol{x}| + \boldsymbol{w}_-^\top \boldsymbol{x}$ (*e.g.*, in 2D one such function is $f(x, y) = x + |y|$). The dual problem for $\overline{R}(f)$ in this instance is given by:

$$\sup_{\substack{\varphi \in C_b(\mathbb{S}^{d-1}\times\mathbb{R}), \boldsymbol{y} \in \mathbb{R}^d \\ \|\mathcal{W}^*\boldsymbol{y} + \mathcal{E}^*\varphi\|_\infty \leq 1}} \boldsymbol{w}_-^\top \boldsymbol{y} + \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \varphi(\boldsymbol{w}, b) d\alpha_+(\boldsymbol{w}, b) \tag{103}$$

Set $\boldsymbol{y}^* = 2\boldsymbol{w}_-^\perp$, and let φ^* be a continuous approximation to $\operatorname{sign}(\alpha_+)$ whose support is localized to an arbitrarily small neighborhood of $\pm(\boldsymbol{w}_+, 0)$. Then the pair $(\boldsymbol{y}^*, \beta^*)$ is feasible since

$$\psi(\boldsymbol{w}, b) := \mathcal{W}^*\boldsymbol{y}^*(\boldsymbol{w}, b) + \mathcal{E}^*\varphi^*(\boldsymbol{w}, b) = \varphi^*(\boldsymbol{w}, b) + \boldsymbol{w}^\top \boldsymbol{w}_- = \begin{cases} 1 & \text{if } \boldsymbol{w} = \pm\boldsymbol{w}_+ \text{ and } b = 0 \\ \boldsymbol{w}^\top \boldsymbol{w}_- & \text{else} \end{cases}$$

and so $|\psi(\boldsymbol{w}, b)| \leq 1$. For these choices of $(\beta^*, \boldsymbol{y}^*)$ the dual objective is $2\|\boldsymbol{w}_-\| + \|f\|_{\mathcal{R}}$, which gives a lower bound on $\overline{R}(f)$. But this is also an upper bound on $\overline{R}(f)$ hence $\overline{R}(f) = \|f\|_{\mathcal{R}} + 2\|\boldsymbol{w}_-\|$. □

F PROPERTIES OF THE $\mathcal{R}$-NORM

Here we prove the properties of $\mathcal{R}$-norm discuseed in Section 4.1, including Proposition 2.

Proposition 12. *The $\mathcal{R}$-norm has the following properties:*

- *(1-homogeneity and triangle inequality) If* $\|f\|_{\mathcal{R}}, \|g\|_{\mathcal{R}} < \infty$*, then* $\|c \cdot f\|_{\mathcal{R}} = |c|\,\|f\|_{\mathcal{R}}$ *for all* $c \in \mathbb{R}$ *and* $\|f+g\|_{\mathcal{R}} \leq \|f\|_{\mathcal{R}} + \|g\|_{\mathcal{R}}$, i.e., $\|\cdot\|_{\mathcal{R}}$ *is a* semi-norm.
- *(Annihilation of affine functions)* $\|f\|_{\mathcal{R}} = 0$ *if and only if f is affine,* i.e., $f(\boldsymbol{x}) = \boldsymbol{v}^\top \boldsymbol{x} + c$ *for some* $\boldsymbol{v} \in \mathbb{R}^d$*,* $c \in \mathbb{R}$*.*
- *(Translation and rotation invariance) If* $g(\boldsymbol{x}) = f(\boldsymbol{U}\boldsymbol{x} + \boldsymbol{y})$ *where* $\boldsymbol{y} \in \mathbb{R}^d$ *and* $\boldsymbol{U} \in \mathbb{R}^{d\times d}$ *is any orthogonal matrix, then* $\|g\|_{\mathcal{R}} = \|f\|_{\mathcal{R}}$*.*
- *(Scaling with dilations/contractions – Proposition 2) Suppose* $\|f\|_{\mathcal{R}} < \infty$*. Let* $f_\varepsilon(\boldsymbol{x}) := f(\boldsymbol{x}/\varepsilon)$*, then* $\|f_\varepsilon\|_{\mathcal{R}} = \varepsilon^{-1}\|f\|_{\mathcal{R}}$*.*

Proof. The 1-homogenity and triangle inequality properties follow immediate from the linearity of all operations and the definition by way of a set supremum.

Clearly $\|f\|_{\mathcal{R}} = 0$ if f is affine. Conversely, suppose $\|f\|_{\mathcal{R}} = 0$ then by the uniqueness in Lemma 8, we have $\alpha = 0$, and so $f = h_{0,\boldsymbol{v},c}$ for some $\boldsymbol{v} \in \mathbb{R}^d$ and $c \in \mathbb{R}$, hence f is affine.

For simplicity we demonstrate proofs of the remaining properties in the case where f is smooth and has sufficient decay so that $\|f\|_{\mathcal{R}} = \gamma_d\|\mathcal{R}\{(-\Delta)^{(d+1)/2}f\}\|_1 = \gamma_d\|\mathcal{L}^{(d+1)/2}\mathcal{R}\{f\}\|_1 < \infty$. We also assume d is odd, so $(-\Delta)^{(d+1)/2}$ is a integer power of the negative Laplacian and $\mathcal{L}^{(d+1)/2}$ is an integer power of the negative partial derivative in b, *i.e.*, $\mathcal{L}^{(d+1)/2} = (-1)^{(d+1)/2}\partial_b^{d+1}$. The general case follows from standard duality arguments.

To show translation invariance, define $f_{(\boldsymbol{y})}(\boldsymbol{x}) := f(\boldsymbol{x} - \boldsymbol{y})$. Then we have $(-\Delta)^{(d+1)/2}f_{(\boldsymbol{y})} = [(-\Delta)^{(d+1)/2}f]_{(\boldsymbol{y})}$. Also, for any function g we see that

$$\mathcal{R}\{g_{(\boldsymbol{y})}\}(\boldsymbol{w}, b) = \mathcal{R}\{g\}(\boldsymbol{w}, b + \boldsymbol{w}^\top\boldsymbol{y}), \tag{104}$$

Therefore,

$$\begin{aligned}
\left\|f_{(y)}\right\|_{\mathcal{R}} &= \int_{\mathbb{S}^{d-1}\times\mathbb{R}} |\mathcal{R}\{(-\Delta)^{(d+1)/2}f_{(\boldsymbol{y})}\}(\boldsymbol{w}, b)|\, d\boldsymbol{w}\, db && (105)\\
&= \int_{\mathbb{S}^{d-1}\times\mathbb{R}} |\mathcal{R}\{(-\Delta)^{(d+1)/2}f\}(\boldsymbol{w}, b + \boldsymbol{w}^\top\boldsymbol{y})|\, d\boldsymbol{w}\, db && (106)\\
&= \int_{\mathbb{S}^{d-1}\times\mathbb{R}} |\mathcal{R}\{(-\Delta)^{(d+1)/2}f\}(\boldsymbol{w}, b)|\, d\boldsymbol{w}\, db = \|f\|_{\mathcal{R}}\,. && (107)
\end{aligned}$$

To show rotation invariance, let $f_{\boldsymbol{U}}(\boldsymbol{x}) = f(\boldsymbol{U}\boldsymbol{x})$ where $\boldsymbol{U}$ is any orthogonal $d \times d$ matrix. Then, using the fact that the Laplacian commutes with rotations, we have $(-\Delta)^{(d+1)/2}f_{\boldsymbol{U}}(\boldsymbol{x}) = (-\Delta)^{(d+1)/2}f(\boldsymbol{U}\boldsymbol{x})$, and since $\mathcal{R}\{g_{\boldsymbol{U}}\}(\boldsymbol{w}, b) = \mathcal{R}\{g\}(\boldsymbol{U}\boldsymbol{w}, b)$, we see that $\mathcal{R}\{(-\Delta)^{(d+1)/2}f_{\boldsymbol{U}}\}(\boldsymbol{w}, b) = \mathcal{R}\{(-\Delta)^{(d+1)/2}f\}(\boldsymbol{U}\boldsymbol{w}, b)$, and so

$$\|f_{\boldsymbol{U}}\|_{\mathcal{R}} = \|f\|_{\mathcal{R}} \tag{108}$$

□

To show the scaling under contractions/dilations (*i.e.*, Proposition 2), let $f_\varepsilon(x) = f(x/\varepsilon)$ for $\varepsilon > 0$. Then

$$\begin{aligned}
\mathcal{R}\{f_\varepsilon\}(\boldsymbol{w}, b) &= \int_{\boldsymbol{w}^\top\boldsymbol{x}=b} f(\boldsymbol{x}/\varepsilon) ds(\boldsymbol{x}) && (109)\\
&= \varepsilon^{d-1}\int_{\boldsymbol{w}^\top\tilde{\boldsymbol{x}}=b/\varepsilon} f(\tilde{\boldsymbol{x}}) ds(\tilde{\boldsymbol{x}}) && (110)\\
&= \varepsilon^{d-1}\mathcal{R}\{f\}(\boldsymbol{w}, b/\varepsilon). && (111)
\end{aligned}$$

Hence, by the Fourier slice theorem we have

$$\begin{aligned}
|\mathcal{L}^{(d+1)/2}\mathcal{R}\{f_\varepsilon\}(\boldsymbol{w}, b)| &= |\partial_b^{d+1}\mathcal{R}\{f_\varepsilon\}(\boldsymbol{w}, b)| && (112)\\
&= \varepsilon^{d-1}\varepsilon^{-d-1}|\partial_b^{d+1}\mathcal{R}\{f\}(\boldsymbol{w}, b/\varepsilon)| && (113)\\
&= \varepsilon^{-2}|\partial_b^{d+1}\mathcal{R}\{f\}(\boldsymbol{w}, b/\varepsilon)| && (114)
\end{aligned}$$

and so

$$\int_{\mathbb{R}\times\mathbb{S}^{d-1}} |\mathcal{L}^{(d+1)/2}\mathcal{R}\{f_\varepsilon\}(w,b)|\,db\,d\boldsymbol{w} = \varepsilon^{-2}\int_{\mathbb{R}\times\mathbb{S}^{d-1}} |(\partial_b^{d+1}\mathcal{R}\{f\}(\boldsymbol{w},b/\varepsilon)|\,db\,d\boldsymbol{w} \tag{115}$$

$$= \varepsilon^{-1}\int_{\mathbb{R}\times\mathbb{S}^{d-1}} |(\partial_b^{d+1}\mathcal{R}\{f\}(\boldsymbol{w},\tilde{b})|\,d\tilde{b}\,d\boldsymbol{w} \tag{116}$$

$$= \varepsilon^{-1}\,\|f\|_{\mathcal{R}} \tag{117}$$

Fourier estimates For any Lipschitz function f we can always interpret Δf in a distributional sense. An interesting special case is when Δf is a distribution of order zero, *i.e.*, a finite measure. In this case, the Fourier transform of Δf, defined as $\widehat{\Delta f}(\boldsymbol{\xi}) := \langle \Delta f, e^{-j2\pi\boldsymbol{x}^\top\boldsymbol{\xi}}\rangle$ for all $\boldsymbol{\xi}\in\mathbb{R}^d$, is a continuous and bounded function, and we can make use of an extension of the Fourier slice theorem to Radon transforms of measures (see, *e.g.*, Boman & Lindskog (2009)) to analyze properties of $\|f\|_{\mathcal{R}}$. In particular, the following result shows that in order for $\|f\|_{\mathcal{R}}$ to be finite, the Fourier transform of Δf (or the Fourier transform of f if it exists classically) must decay at a certain rate.

Proposition 13. *Suppose Δf is a finite measure. Then $\|f\|_{\mathcal{R}}$ is finite only if $\widehat{\Delta f}(\sigma\cdot\boldsymbol{w}) = O(|\sigma|^{-(d-1)})$ as $|\sigma|\to\infty$ for all $\boldsymbol{w}\in\mathbb{S}^{d-1}$. If additionally $f\in L^1(\mathbb{R}^d)$, then $\|f\|_{\mathcal{R}}$ is finite only if $\widehat{f}(\sigma w) = O(|\sigma|^{-(d+1)})$ as $|\sigma|\to\infty$ for all $\boldsymbol{w}\in\mathbb{S}^{d-1}$.*

Proof. If $\Delta f\in M(\mathbb{R}^d)$ is a finite measure then its Radon transform $\mathcal{R}\{\Delta f\}\in M(\mathbb{P}^d)$ exists as a finite measure, *i.e.*, we can define $\mathcal{R}\{\Delta f\}$ via duality as $\langle\mathcal{R}\{\Delta f\},\varphi\rangle = \langle\Delta f,\mathcal{R}^*\{\varphi\}\rangle$ for all $\varphi\in C_0(\mathbb{R}^d)$ (see, *e.g.*, Boman & Lindskog (2009)). Additionally, the restriction $\mathcal{R}\{\Delta f\}(w,\cdot)\in M(\mathbb{R})$ is well-defined finite measure for all $\boldsymbol{w}\in\mathbb{S}^{d-1}$, and its 1D Fourier transform in the b variable is given by

$$\mathcal{F}_b\mathcal{R}\{\Delta f\}(\boldsymbol{w},\sigma) = \widehat{\Delta f}(\sigma\cdot\boldsymbol{w}) \text{ for all } \boldsymbol{w}\in\mathbb{S}^{d-1},\sigma\in\mathbb{R}. \tag{118}$$

By Lemma 8, $\|f\|_{\mathcal{R}}$ is finite if and only if the functional $L_f(\psi) = -\gamma_d\langle f,(-\Delta)^{(d+1)/2}\mathcal{R}^*\{\psi\}\rangle$ defined for all $\psi\in\mathcal{S}(\mathbb{P}^d)$ extends to a unique measure $\alpha\in M(\mathbb{P}^d)$. We compute its Fourier transform of α in the b variable via duality: for all $\varphi\in\mathcal{S}(\mathbb{P}^d)$ we have

$$\langle\mathcal{F}_b\alpha,\varphi\rangle = \langle\alpha,\mathcal{F}_b\varphi\rangle \tag{119}$$

$$= \gamma_d\langle f,(-\Delta)^{(d+1)/2}\mathcal{R}^*\{\mathcal{F}_b\varphi\}\rangle \tag{120}$$

$$= -\gamma_d\langle\Delta f,(-\Delta)^{(d-1)/2}\mathcal{R}^*\{\mathcal{F}_b\varphi\}\rangle \tag{121}$$

$$= -\gamma_d\langle\Delta f,\mathcal{R}^*\{(-\mathcal{L})^{(d-1)/2}\mathcal{F}_b\varphi\}\rangle \tag{122}$$

$$= -\gamma_d\langle\Delta f,\mathcal{R}^*\{\mathcal{F}_b(|\sigma|^{d-1}\varphi)\}\rangle \tag{123}$$

$$= -\gamma_d\langle\mathcal{R}\{\Delta f\},\mathcal{F}_b(|\sigma|^{d-1}\varphi)\rangle \tag{124}$$

$$= -\gamma_d\langle\mathcal{F}_b\mathcal{R}\{\Delta f\},|\sigma|^{d-1}\varphi\rangle \tag{125}$$

$$= -\gamma_d\langle|\sigma|^{d-1}\mathcal{F}_b\mathcal{R}\{\Delta f\},\varphi\rangle \tag{126}$$

Since $\mathcal{F}_b\mathcal{R}\{\Delta f\}$ is defined pointwise for all $(\boldsymbol{w},b)\in\mathbb{S}^{d-1}\times\mathbb{R}$ so is $\mathcal{F}_b\alpha$ and we have

$$(\mathcal{F}_b\alpha)(\boldsymbol{w},\sigma) = -\gamma_d|\sigma|^{d-1}\mathcal{F}_b\mathcal{R}\{\Delta f\}(\boldsymbol{w},\sigma) = -\gamma_d|\sigma|^{d-1}\widehat{\Delta f}(\sigma\cdot\boldsymbol{w}). \tag{127}$$

Finally, since α is a finite measure, we know $\|\mathcal{F}_b\alpha\|_\infty \le \|\alpha\|_1 = O(1)$, which gives the first result. If additionally $f\in L^1(\mathbb{R}^d)$ then we have $\widehat{\Delta f}(\boldsymbol{\xi}) = \|\boldsymbol{\xi}\|^2\widehat{f}(\boldsymbol{\xi})$, and so $(\mathcal{F}_b\alpha)(\boldsymbol{w},b) = |\sigma|^{d+1}\widehat{\Delta f}(\sigma\cdot\boldsymbol{w})$ which gives the second result. □

G Upper and Lower bounds

Here we prove several upper and lower bounds for the $\mathcal{R}$-norm. Proposition 3 is an immediate corollary of the following upper bound:

Proposition 14. *If $(-\Delta)^{(d+1)/2}f$ is a finite measure, then*

$$\|f\|_{\mathcal{R}} \le \gamma_d c_d\|(-\Delta)^{(d+1)/2}f\|_1, \tag{128}$$

In particular, if $(-\Delta)^{(d+1)/2}f$ exists in a weak sense then $\|\cdot\|_1$ can be interpreted as the L^1-norm.

Proof. Straight from definitions we have

$$\|f\|_{\mathcal{R}} = \sup\Big\{\gamma_d\langle f, (-\Delta)^{(d+1)/2}\mathcal{R}^*\{\psi\}\rangle : \psi \in \mathcal{S}(\mathbb{P}^d), \|\psi\|_\infty \leq 1\Big\} \tag{129}$$

$$= \sup\Big\{\gamma_d\langle (-\Delta)^{(d+1)/2} f, \mathcal{R}^*\{\psi\}\rangle : \psi \in \mathcal{S}(\mathbb{P}^d), \|\psi\|_\infty \leq 1\Big\} \tag{130}$$

$$\leq \sup\Big\{\gamma_d\langle (-\Delta)^{(d+1)/2} f, \varphi\rangle : \varphi \in C_0(\mathbb{R}^d), \|\varphi\|_\infty \leq c_d\Big\} \tag{131}$$

$$= \gamma_d c_d \|(-\Delta)^{(d+1)/2} f\|_1 \tag{132}$$

where we used the fact that $\mathcal{R}^*\{\varphi\} \in C_0(\mathbb{R}^d)$ for $\varphi \in \mathcal{S}(\mathbb{P}^d)$ (Solmon, 1987, Corollary 3.6) and we have $\|\mathcal{R}^*\{\varphi\}\|_\infty \leq c_d$ for all $\varphi \in \mathcal{S}(\mathbb{P}^d)$ such that $\|\varphi\|_\infty \leq 1$ since

$$|\mathcal{R}^*\{\varphi\}(\boldsymbol{x})| \leq \int_{\mathbb{S}^{d-1}} |\varphi(\boldsymbol{w}, \boldsymbol{w}^\top \boldsymbol{x})|\, d\boldsymbol{w} \leq \int_{\mathbb{S}^{d-1}} d\boldsymbol{w} = c_d. \tag{133}$$

□

The following result also gives a useful lower bound on the $\mathcal{R}$-norm.

Proposition 15. *If $f \in \mathrm{Lip}(\mathbb{R}^d)$ then*

$$\|f\|_{\mathcal{R}} \geq \sup\big\{\langle f, \Delta\varphi\rangle : \varphi \in \mathcal{S}(\mathbb{R}^d), \|\mathcal{R}\{\varphi\}\|_\infty \leq 1\big\}. \tag{134}$$

Proof. Let $\mathcal{S}_H(\mathbb{P}^d) \subset \mathcal{S}(\mathbb{P}^d)$ denote the image of $\mathcal{S}(\mathbb{R}^d)$ under the Radon transform. Then

$$\|f\|_{\mathcal{R}} = \sup\Big\{\gamma_d\langle f, (-\Delta)^{(d+1)/2}\mathcal{R}^*\{\psi\}\rangle : \psi \in \mathcal{S}(\mathbb{P}^d), \|\psi\|_\infty \leq 1\Big\} \tag{135}$$

$$\geq \sup\Big\{\gamma_d\langle f, (-\Delta)^{(d+1)/2}\mathcal{R}^*\{\psi\}\rangle : \psi \in \mathcal{S}_H(\mathbb{P}^d), \|\psi\|_\infty \leq 1\Big\} \tag{136}$$

$$= \sup\Big\{\gamma_d\langle f, (-\Delta)^{(d+1)/2}\mathcal{R}^*\{\mathcal{R}\{\varphi\}\}\rangle : \varphi \in \mathcal{S}(\mathbb{R}^d), \|\mathcal{R}\{\varphi\}\|_\infty \leq 1\Big\} \tag{137}$$

$$= \sup\big\{\langle f, \Delta\varphi\rangle : \varphi \in \mathcal{S}(\mathbb{R}^d), \|\mathcal{R}\{\varphi\}\|_\infty \leq 1\big\} \tag{138}$$

where in the last step we used the inversion formula: $\varphi = \gamma_d(-\Delta)^{(d-1)/2}\mathcal{R}^*\{\mathcal{R}\{\varphi\}\}$ for all $\varphi \in \mathcal{S}(\mathbb{R}^d)$. □

Furhter simplifying the lower bound above gives the following.

Proposition 16. *If $f \in \mathrm{Lip}(\mathbb{R}^d)$ then*

$$\|f\|_{\mathcal{R}} \geq \sup\big\{\langle f, \Delta\varphi\rangle : \varphi \in \mathcal{S}(\mathbb{R}^d), \|\varphi\|_1 \leq 1\big\}. \tag{139}$$

In particular, if Δf exists in a weak sense then $\|f\|_{\mathcal{R}} \geq \|\Delta f\|_\infty$.

Proof. If $\|\varphi\|_1 = \int |\varphi(\boldsymbol{x})|\, d\boldsymbol{x} \leq 1$ then clearly $|\mathcal{R}\{\varphi\}(\boldsymbol{w}, b)| = |\int_{\boldsymbol{w}^\top \boldsymbol{x}=b} \varphi(\boldsymbol{x})d\boldsymbol{x}| \leq \int_{\boldsymbol{w}^\top \boldsymbol{x}=b} |\varphi(\boldsymbol{x})|\, d\boldsymbol{x} \leq 1$. Hence $\|\varphi\|_1 \leq 1$ implies $\|\mathcal{R}\{\varphi\}\|_\infty \leq 1$. Combining this with the previous proposition gives the first bound. Additionally, by the dual definition of the L^∞ norm, and since $\mathcal{S}(\mathbb{R}^d)$ is dense in $L^1(\mathbb{R}^d)$, the second bound follows. □

H RADIAL BUMP FUNCTIONS

Proof of Proposition 4. Assume $f \in L^1(\mathbb{R}^d)$ so that its Radon transform $\mathcal{R}\{f\}$ is well-defined, and for simplicity assume d is odd. Note that for a radially symmetric function we have $\mathcal{R}\{f\}(\boldsymbol{w}, b) = \rho(b)$ for some even function $\rho \in L^1(\mathbb{R})$, *i.e.*, the Radon transform of a radially symmetric function does not depend on the unit direction $\boldsymbol{w} \in \mathbb{S}^{d-1}$. Supposing $\partial^{(d+1)}\rho(b)$ exists either as a function or a measure, we have

$$\|f\|_{\mathcal{R}} = \gamma_d\|\partial_b^{d+1}\mathcal{R}\{f\}\|_1 = \gamma_d c_d \int |\partial^{d+1}\rho(b)| db, \tag{140}$$

where $c_d = \int_{\mathbb{S}^{d-1}} d\boldsymbol{w} = \frac{2\pi^{d/2}}{\Gamma(d/2)}$.

Now we derive an expression for $\rho(b)$ in terms of g. First, since $\rho(b) = \mathcal{R}\{f\}(\boldsymbol{w}, b)$ for any $\boldsymbol{w} \in \mathbb{S}^{d-1}$, we can choose $\boldsymbol{w} = \boldsymbol{e}_1 = (1, 0, ..., 0)$, which gives

$$\rho(b) = \mathcal{R}\{f\}(\boldsymbol{e}_1, b) = \int_{x_1=b} g(\|\boldsymbol{x}\|)dx_2 \cdots dx_d = \int_{\mathbb{R}^{d-1}} g(\sqrt{b^2 + \|\tilde{\boldsymbol{x}}\|^2})d\tilde{\boldsymbol{x}} \tag{141}$$

where we have set $\tilde{\boldsymbol{x}} = (x_2, ..., x_d)$. Changing to polar coordinates over $\mathbb{R}^{d-1}$, we have

$$\rho(b) = \int_{\mathbb{R}^{d-1}} g(\sqrt{b^2 + \|\tilde{\boldsymbol{x}}\|^2})d\tilde{\boldsymbol{x}} = c_{d-1} \int_0^\infty g(\sqrt{b^2 + r^2})r^{d-2}dr. \tag{142}$$

By the change of variables $t^2 = b^2 + r^2$, $t > 0$, we have

$$\rho(b) = c_{d-1} \int_b^\infty g(t)(t^2 - b^2)^{(d-3)/2}t\, dt. \tag{143}$$

Hence, we see that

$$\|f\|_{\mathcal{R}} = \frac{1}{(d-2)!} \left\| \partial_b^{(d+1)} \left[\int_b^\infty g(t)(t^2 - b^2)^{(d-3)/2}t\, dt \right] \right\|_1 \tag{144}$$

where we used the fact that $\gamma_d c_d c_{d-1} = \frac{1}{(d-2)!}$.

Calculations in Example 3. Let $f(\boldsymbol{x}) = g_{d,k}(\|\boldsymbol{x}\|)$ with $\boldsymbol{x} \in \mathbb{R}^d$ where

$$g_{d,k}(r) = \begin{cases} (1 - r^2)^k & \text{if } 0 \leq r < 1 \\ 0 & \text{if } r \geq 1. \end{cases} \tag{145}$$

for any $k > 0$. Then a straightforward calculation using (143) gives

$$\rho(b) = \begin{cases} C_{d,k}(1 - b^2)^{k + \frac{d-1}{2}} & \text{if } |b| < 1 \\ 0 & \text{if } b \geq 1. \end{cases} \tag{146}$$

where $C_{d,k} = \frac{\Gamma((d-3)/2) \cdot \Gamma(1+k)}{2\Gamma((d+1)/2)+k)}$. Hence, we have $\|f\|_{\mathcal{R}}$ finite if and only if $\partial_b^d \rho(b)$ has bounded variation, which is true if and only if $k - d + \frac{d-1}{2} \geq 0$, or equivalently, $k \geq \frac{d+1}{2}$. For example, if $d = 3$ then we need $k \geq 2$ in order for $\|f\|_{\mathcal{R}}$ to be finite, consistent with the previous example.

To illustrate scaling of $\|f\|_{\mathcal{R}}$ with dimension d, we set $k = (d+1)/2 + 2 = (d+5)/2$ so that $\rho(b) = C_{d,(d+5)/2}(1 - b^2)^{d+2}$ for $|b| \leq 1$ and $\rho(b) = 0$ otherwise. Then we can show that $|\partial^{d+1}\rho(b)| \leq |\partial^{d+1}\rho(0)|$ for $|b| \leq 1$ and $\partial^{d+1}\rho(b) = 0$ for all $|b| \geq 1$. Therefore,

$$\|f\|_{\mathcal{R}} = \frac{1}{(d-2)!} \int_{-1}^1 |\partial^{d+1}\rho(b)| \leq \frac{2}{(d-2)!}|\partial^{d+1}\rho(0)| \tag{147}$$

Performing a binomial expansion of $\rho(b)$ and taking derivatives, we obtain

$$\frac{2}{(d-2)!}|\partial^{d+1}\rho(0)| = 2C_{d,(d+5)/2} \binom{d+2}{(d+1)/2} (d+1)d(d-1) = 2d(d+5) \tag{148}$$

for all odd $d \geq 3$. By the lower bound in Proposition 16, we also have $\|f\|_{\mathcal{R}} \geq \|\Delta f\|_\infty = |\Delta f(\mathbf{0})| = d(d+5)$. Hence $\|f\|_{\mathcal{R}} = \Theta(d^2)$.

I Piecewise Linear Functions

Proof of Proposition 5

Proof. Assume f is continuous and piecewise linear with compact support. Then the distributional Laplacian Δf decomposes into a sum of Dirac measures supported on restricted hyperplanes H_i defining boundaries between piecewise linear regions, *i.e.*, we have

$$\langle \Delta f, \varphi \rangle = \sum_{i=1}^n c_i \oint_{H_i} \varphi(\boldsymbol{x})\, ds(\boldsymbol{x}). \tag{149}$$

for some coefficients $c_i \in \mathbb{R}$, where ds indicates the $d-1$ dimensional surface measure on H_i. Hence, Δf is a distribution of order-0, *i.e.*, it can be identified with a measure having finite total variation and its Fourier transform is given by

$$\widehat{\Delta f}(\boldsymbol{\xi}) = \sum_{i=1}^{n} c_i \int_{H_i} e^{j2\pi\xi^\top \boldsymbol{x}}\, ds(\boldsymbol{x}). \tag{150}$$

Without loss of generality, assume there is at least one H_i whose normal vector $\boldsymbol{w} \in \mathbb{S}^{d-1}$, *i.e.*, $\boldsymbol{w}^\top \boldsymbol{x} = 0$ for all $\boldsymbol{x} \in H_i$, is distinct (up to sign) from the normal vectors of all the other H_i (if this is not the case, then we can perturb the function slightly and obtain the same result). If we let $\boldsymbol{\xi} = \sigma\, \boldsymbol{w}$ for any $\sigma \in \mathbb{R}$, then the integral $\int_{H_i} e^{j2\pi\boldsymbol{\xi}^\top \boldsymbol{x}} ds(\boldsymbol{x}) = \oint_{H_i} ds(\boldsymbol{x}) = s(H_i)$, where $s(H_i)$ the is $(d-1)$-dimension surface measure of H_i, and in particular it is $\Theta(1)$ (upper and lower bounded by a constant) as $|\sigma| \to \infty$, while the remainder of the integrals are at least $O(|\sigma|^{-1})$, since they reduce to a product of d sinc functions, at least one of which is non-constant.

Hence for all $\sigma \in \mathbb{R}$ we have $\widehat{\Delta f}(\sigma \boldsymbol{w}) = \Theta(1)$, and by Proposition 13 this implies $\|f\|_{\mathcal{R}}$ must be infinite. □