A Gap Between the Gaussian RKHS and Neural Networks: An Infinite-Center Asymptotic Analysis

Akash Kumar AKK002@UCSD.EDU
Department of Computer Science and Engineering
University of California, San Diego

Rahul Parhi RAHUL@UCSD.EDU
Department of Electrical and Computer Engineering
University of California, San Diego

Mikhail Belkin MBELKIN@UCSD.EDU
Department of Computer Science and Engineering
Halıcıoğlu Data Science Institute
University of California, San Diego

Editors: Nika Haghtalab and Ankur Moitra

Abstract

Recent works have characterized the function-space inductive bias of infinite-width bounded-norm single-hidden-layer neural networks as a kind of bounded-variation-type space. This novel neural network Banach space encompasses many classical multivariate function spaces, including certain Sobolev spaces and the spectral Barron spaces. Notably, this Banach space also includes functions that exhibit less classical regularity, such as those that only vary in a few directions. On bounded domains, it is well-established that the Gaussian reproducing kernel Hilbert space (RKHS) strictly embeds into this Banach space, demonstrating a clear gap between the Gaussian RKHS and the neural network Banach space. It turns out that when investigating these spaces on unbounded domains, e.g., all of $\mathbb{R}^d$, the story is fundamentally different. We establish the following fundamental result: Certain functions that lie in the Gaussian RKHS have infinite norm in the neural network Banach space. This provides a nontrivial gap between kernel methods and neural networks by exhibiting functions that kernel methods easily represent, whereas neural networks cannot.

1. Introduction

In supervised learning, we observe samples with corresponding labels, which may represent classes or continuous values. Our primary objective is to construct a function $f : \mathbb{R}^d \to \mathbb{R}$ based on these observations that can accurately predict labels for new, unseen data points. Traditionally, reproducing kernel Hilbert spaces (RKHS) have provided a principled framework for this task, offering both theoretical guarantees and practical algorithms. Their power stems from the representer theorem, which ensures that optimal solutions can be expressed as combinations of kernel functions centered at the training points.

However, the landscape of machine learning has evolved significantly with the emergence of neural networks, which have demonstrated remarkable success across diverse applications over kernel methods. The simplest neural architecture—the single-hidden layer network—builds upon the concept of ridge functions, which map $\mathbb{R}^d \to \mathbb{R}$ via the form $\boldsymbol{x} \mapsto \sigma(\boldsymbol{w}^\top \boldsymbol{x})$, where $\sigma : \mathbb{R} \to \mathbb{R}$

is a univariate function and $\boldsymbol{w} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$. In practice, these networks combine multiple ridge functions:

$$\boldsymbol{x} \mapsto \sum_{k=1}^{K} v_k \, \sigma(\boldsymbol{w}_k^{\mathsf{T}} \boldsymbol{x} - b_k), \tag{1}$$

where K represents the network width, $v_k \in \mathbb{R}$ and $\boldsymbol{w}_k \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ are weights, and $b_k \in \mathbb{R}$ are biases. While RKHS methods suffer from the curse of dimensionality, neural networks can overcome it by learning effective low-dimensional representations (Ghorbani et al., 2021b; von Luxburg and Bousquet, 2004).

A fundamental question is to compare the approximation capabilities of neural networks with those of RKHS corresponding to different kernels. For example, Mei et al. (2016) showed that if the target function is a single neuron, neural networks can learn efficiently using roughly $d \log d$ samples, whereas the corresponding RKHS requires a sample size that grows polynomially in the dimension d (see also Yehudai and Shamir (2019); Ghorbani et al. (2019)).

Recent work (Parhi and Nowak, 2021, 2023a) has studied the Banach-space optimality of single-hidden-layer (shallow ReLU) networks over both bounded and unbounded domains $\Omega \subseteq \mathbb{R}^d$. There, the authors established a representer theorem which demonstrates that solutions to data-fitting problems in these networks naturally reside in a kind of bounded variation space, referred to as the second-order Radon bounded variation space $\mathcal{R}\mathrm{BV}^2(\Omega)$. These spaces, in turn, contain several classical multivariate function spaces, including certain Sobolev spaces as well as certain spectral Barron spaces (Barron, 1993). For instance, Parhi and Nowak (2023a) have shown that the Sobolev space $H^{d+1}(\Omega)$ embeds into $\mathcal{R}\mathrm{BV}^2(\Omega)$ for any bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$. Moreover, on any bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, the Gaussian reproducing kernel Hilbert space $\mathcal{H}^{\mathsf{Gauss}}(\Omega)$ is known to embed into the Sobolev space $H^s(\Omega)$ for all $s > 0$ (see Corollary 4.36 of Steinwart and Christmann (2008)). This observation appears to highlight limitations of Gaussian kernel machines when compared to neural networks on *bounded domains*. Consequently, a natural question arises.

Are Gaussian kernel machines restrictive in approximating general functions?

Conversely, one may also ask the following question.

To what extent can we approximate functions using shallow neural networks?

To that end, Ghorbani et al. (2021a) demonstrated that the gap between neural network approximations and kernel methods can be narrowed when the intrinsic dimensionality of the target function is well captured by the covariates of the data. In this paper, we take a different perspective: While the Gaussian RKHS may seem rather limited in a bounded domain, we show that on unbounded domains, in particular, on $\mathbb{R}^d$ with fixed dimension d, there exist functions in $\mathcal{H}^{\mathsf{Gauss}}(\mathbb{R}^d)$ with unbounded $\mathcal{R}\mathrm{BV}^2(\mathbb{R}^d)$-norm.

The key idea behind our analysis is that, in the regime of kernel machines with infinite centers on $\mathbb{R}^d$, there exist functions of the form $f = \sum_{i=1}^{\infty} \alpha_i k(\boldsymbol{x}_i, \cdot)$ with bounded RKHS norm, but the infinite sequence $\{\alpha_i\}$ has an *unbounded* ℓ_1-norm (see Example 2 in Section 3). This fact can be exploited to design a sequence of functions $\{f_n\}$ whose $\mathcal{R}\mathrm{BV}^2(\mathbb{R}^d)$-norm is diverging as $n \to \infty$ (see Theorem 7 in Section 5). An important step in this study is we compute an explicit form for $\mathcal{R}\mathrm{BV}^2(\mathbb{R}^d)$-norm of a Gaussian kernel machine, and further simplify the form using well-known Hermite polynomials. This form provides an interpretable characterization of these kernel machines, which is of independent interest for future studies.

2. Related Work

Approximability with Kernel Methods Bach (2017) studied various classes of single-/multi-index models with low intrinsic dimension and bounded $\mathcal{R}\mathrm{BV}^2(\mathbb{R}^d)$-norm. In contrast, Ghorbani et al. (2019) showed that if the covariates have the same dimension as the low intrinsic dimension of the target function, kernel and neural network approximations can be competitive. Empirically, some works show that the curse of dimensionality with kernel methods can be handled with an appropriate choice of dataset-specific kernels (Arora et al., 2019; Novak et al., 2018; Shankar et al., 2020) or mirroring neural network training dynamics closely to kernel methods (Mei et al., 2018; Sirignano and Spiliopoulos, 2020; Rotskoff and Vanden-Eijnden, 2022; Chizat and Bach, 2018). Furthermore, Petrini et al. (2023) showed that compared to a network that learns sparse representations while the target function is constant or smooth along certain directions of the input space, lazy training (via random feature kernel or the NTK) yields better performance. But a wide body of work has also shown a gap in approximation with neural networks capturing a richer and more nuanced class of functions compared to kernel methods (see (Allen-Zhu and Li, 2019; Mei et al., 2018; Yehudai and Shamir, 2019; Ghorbani et al., 2019)). In our work, we show that while Gaussian RKHS is embedded within neural networks in bounded domains, in the unbounded regime there exists a non-trivial gap between $\mathcal{H}^{\mathsf{Gauss}}(\mathbb{R}^d)$ and $\mathcal{R}\mathrm{BV}^2(\mathbb{R}^d)$.

Function Spaces of Shallow Networks The function space $\mathcal{R}\mathrm{BV}^2(\Omega)$ naturally characterizes the function approximation and representation capabilities of shallow ReLU neural networks (Ongie et al., 2020). Parhi and Nowak (2021) established a *representer theorem*, showing that solutions to variational problems over $\mathcal{R}\mathrm{BV}^2(\Omega)$ correspond to single-hidden layer ReLU networks with weight decay regularization. Unlike RKHSs $\mathcal{R}\mathrm{BV}^2(\Omega)$ can efficiently represent functions with a low-dimensional structure. Moreover, neural networks trained with weight decay achieve near-minimax optimal estimation rates for functions in $\mathcal{R}\mathrm{BV}^2(\Omega)$, while kernel methods provably cannot (Parhi and Nowak, 2023a). This suggests that on bounded domains, RKHSs are quite restrictive, while $\mathcal{R}\mathrm{BV}^2(\Omega)$ provides a more expressive framework. For further details see (Ongie et al., 2020; Parhi and Nowak, 2021, 2022, 2023a,b; Bartolucci et al., 2023; Parhi and Unser, 2025)

Embeddings of RKHSs and $\mathcal{R}\mathrm{BV}^2(\Omega)$ For any bounded Lipshitz domain $\Omega \subseteq \mathbb{R}^d$, it is well-known that the Sobolev space $H^s(\Omega)$ is (equivalent to) an RKHS if and only if $s > d/2$. For example, the Laplace and Matérn kernels are associated with Sobolev RKHSs (see, e.g., Kanagawa et al., 2018, Example 2.6). In contrast, Zhou (2003) and Steinwart and Christmann (cf., 2008, Corollary 4.36) showed that the Gaussian RKHS $\mathcal{H}^{\mathsf{Gauss}}(\Omega)$ is contained in $\mathcal{H}^{\mathsf{Gauss}}(\Omega) \subset H^s(\Omega)$ for all $s \geq 0$. Recent work has further demonstrated that the RKHSs of typical neural tangent kernel (NTK) and neural network Gaussian process (NNGP) kernels for the ReLU activation function are equivalent to the Sobolev spaces $H^{(d+1)/2}(\mathbb{S}^d)$ and $H^{(d+3)/2}(\mathbb{S}^d)$, respectively (Bietti and Bach, 2021; Chen and Xu, 2021). Steinwart et al. (2009) has shown that an optimal learning rates in Sobolev RKHSs can be achieved by cross-validating the regularization parameter. On another front, embedding properties relating Sobolev spaces and the second-order Radond-domain bounded variation space has been explored. For example, Ongie et al. (2020) showed that $W^{d+1}(L_1(\mathbb{R}^d))$ embeds in $\mathcal{R}\mathrm{BV}^2(\mathbb{R}^d)$. More recently, Mao et al. (2024) established a sharp bound by proving that $W^s(L_p(\Omega))$ with $s \geq 2 + (d+1)/2$ for $p \geq 2$ embeds in $\mathcal{R}\mathrm{BV}^2(\Omega)$ for bounded domains $\Omega \subset \mathbb{R}^d$.

3. Problem Setup and Preliminaries

3.1. Gaussian Reproducing Kernel Hilbert Space

We begin by defining a reproducing kernel Hilbert space (RKHS) associated with a Gaussian kernel on an infinite domain. For a given positive definite Mahalanobis matrix $\mathbf{M} \in \mathrm{Sym}_+(\mathbb{R}^{d\times d})$, we define the Gaussian kernel $k_{\mathbf{M}} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as

$$k_{\mathbf{M}}(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|_{\mathbf{M}}^2}{2\sigma^2}\right), \tag{2}$$

where $\sigma > 0$ is a fixed scale parameter and the Mahalanobis distance is defined as

$$\|\boldsymbol{x}-\boldsymbol{y}\|_{\mathbf{M}}^2 = (\boldsymbol{x}-\boldsymbol{y})^{\mathsf{T}}\mathbf{M}(\boldsymbol{x}-\boldsymbol{y}). \tag{3}$$

The corresponding RKHS $\mathcal{H}$ is defined as the closure[1] of the linear span of kernel functions:

$$\mathcal{H} := \mathsf{cl}\left(\left\{ f : \mathcal{X} \to \mathbb{R} \,\middle|\, n \in \mathbb{N}, f(\cdot) = \sum_{i=1}^{n} \alpha_i \cdot k_{\mathbf{M}}(\boldsymbol{x}_i, \cdot),\, \boldsymbol{x}_i \in \mathbb{R}^d \right\}\right), \tag{4}$$

where the (squared) RKHS norm $\|\cdot\|_{\mathcal{H}}^2$ of a kernel machine $f \in \mathcal{H}$ is defined as $\|f\|_{\mathcal{H}}^2 = \sum_{i,j} \alpha_i \alpha_j k_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Alternately, we can write $\|f\|_{\mathcal{H}}^2 = \alpha^{\mathsf{T}}\mathbf{K}\alpha$ where $\mathbf{K} = (k_{\mathbf{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j}$ is an $n \times n$ matrix.

3.2. Separated Sets and Function Spaces

For our analysis, we introduce two key definitions of separated sets that play a crucial role in our theoretical development.

Definition 1 ($(\boldsymbol{\beta}, \delta)$-separated set) *For any given scalar $\delta > 0$ and a vector $\boldsymbol{\beta} \in \mathbb{R}^d$, a $(\boldsymbol{\beta}, \delta)$-separated subset of size $n \in \mathbb{N}$ is defined as*

$$\mathcal{C}_n(\boldsymbol{\beta}, \delta) := \left\{ \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \,\middle|\, \forall\, i, j,\ |\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{x}_i - \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{x}_j| \geq \delta \right\}. \tag{5}$$

This could be further generalized to the notion

Definition 2 ($(\boldsymbol{\beta}, \delta, \eta)$-separated set) *For any given scalars $\delta, \eta > 0$ and a vector $\boldsymbol{\beta} \in \mathbb{R}^d$ a $(\boldsymbol{\beta}, \delta, \eta)$-separated subset of size $n \in \mathbb{N}$ is defined as*

$$\mathcal{C}_n(\boldsymbol{\beta}, \delta, \eta) := \left\{ \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \,\middle|\, \forall\, i, j, \boldsymbol{\beta}' \in \mathbb{R}^d \text{ s.t. } \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\beta}' \geq \eta \|\boldsymbol{\beta}\| \left\|\boldsymbol{\beta}'\right\|, |\boldsymbol{\beta}'^{\mathsf{T}}\boldsymbol{x}_i - \boldsymbol{\beta}'^{\mathsf{T}}\boldsymbol{x}_j| \geq \delta \right\}. \tag{6}$$

Example 1 *Let $\boldsymbol{\beta} = (1, 0, \ldots, 0)$. For all $\eta_0 \geq \eta$, pick $\boldsymbol{\beta}' = \left(\eta_0, \sqrt{1-\eta_0^2}, 0, \ldots, 0\right)$ so that $\|\boldsymbol{\beta}'\| = 1$ and $\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\beta}' = \eta_0 \geq \eta$. Now define*

$$\boldsymbol{x}_i := (i-1)\,\delta\,\boldsymbol{\beta}', \quad i = 1, \ldots, n. \tag{7}$$

For $i \neq j$, we have

$$\left|\boldsymbol{\beta}'^{\mathsf{T}}\boldsymbol{x}_i - \boldsymbol{\beta}'^{\mathsf{T}}\boldsymbol{x}_j\right| = |(i-j)|\,\delta\,\|\boldsymbol{\beta}'\|^2 = |i-j|\,\delta \geq \delta. \tag{8}$$

Hence, $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is in $(\boldsymbol{\beta}, \delta, \eta)$-separated subset of size n.

1. With respect to the norm topology on $\mathcal{H}$.

Definition 3 (Unbounded Combinations) *For a kernel machine $f \in \mathcal{H}$ with representation $f = \sum_{i=1}^{\infty} \alpha_i k(\boldsymbol{x}_i, \cdot)$, we say the coefficient vector $\alpha = (\alpha_i)_{i=1}^{\infty}$ is unbounded with respect to f if $\|\alpha\|_{\ell_1} = \sum_{i=1}^{\infty} |\alpha_i| = \infty$.*

Example 2 *Consider a kernel machine $f \in \mathcal{H}$ corresponding to the combination $\alpha = (a_n)$ defined by $a_n = \frac{1}{n}$ for each $n \in \mathbb{N}$ and a sequence of centers $(\boldsymbol{x}_n) \subset \mathbb{R}^d$ such that for all i, j, $\|\boldsymbol{x}_i - \boldsymbol{x}_j\| \geq |i - j|\delta$ for some fixed scalar $\delta > 0$. For this construction, Gaussian RKHS norm $\alpha^{\mathsf{T}}\mathbf{K}\alpha = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \alpha_i \alpha_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) < \infty$, but $\|\alpha\|_{\ell_1}$ is unbounded.*

We provide the proof of the statement in the example above in Appendix D.

3.3. Probabilist's Hermite Polynomials

Probabilist's Hermite polynomials (Szegő, 1975), denoted by $\mathrm{He}_d(z) : \mathbb{R} \to \mathbb{R}$, are defined by the generating function

$$\exp\Big(zt - \frac{t^2}{2}\Big) = \sum_{d=0}^{\infty} \mathrm{He}_d(z) \frac{t^d}{d!}, \tag{9}$$

and they are orthogonal with respect to the standard normal density

$$\int_{-\infty}^{\infty} \mathrm{He}_d(z)\, \mathrm{He}_{d'}(z)\, \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{z^2}{2}\Big)\, dz = d!\, \delta_{dd'}, \tag{10}$$

where $\delta_{dd'}$ is the Kronecker delta. We use the notation H_d to denote the polynomial unless stated otherwise.

3.4. Radon Transform and the Second-Order Radon-Domain Bounded Variation Space

Radon Transform For a function $f : \mathbb{R}^d \to \mathbb{R}$, its Radon transform $\mathcal{R}\{f\}$ is defined by

$$\mathcal{R}\{f\}(\boldsymbol{\beta}, t) = \int_{\{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{x} = t\}} f(\boldsymbol{x})\, ds(\boldsymbol{x}), \tag{11}$$

where $\boldsymbol{u} \in \mathbb{S}^{d-1}$, $t \in \mathbb{R}$, and $ds(\boldsymbol{x})$ is the $(d-1)$-dimensional Lebesgue measure on the hyperplane

Radon Bounded Variation Space We define the second-order Radon bounded variation space $\mathcal{R}\mathrm{BV}^2(\mathbb{R}^d)$ as:

$$\mathcal{R}\mathrm{BV}^2(\mathbb{R}^d) = \left\{ f : \mathbb{R}^d \to \mathbb{R} \text{ is measurable} : \begin{array}{c} \mathcal{R}\mathrm{TV}^2(f) < \infty, \\ \operatorname*{ess\,sup}_{\boldsymbol{x} \in \mathbb{R}^d} |f(\boldsymbol{x})|(1 + \|\boldsymbol{x}\|)^{-1} < \infty \end{array} \right\}, \tag{12}$$

where the second-order Radon total variation norm $\mathcal{R}\mathrm{TV}^2(f)$ is a seminorm defined by

$$\mathcal{R}\mathrm{TV}^2(f) = c_d \|\partial_t^2 \Lambda^{d-1} \mathcal{R} f\|_{\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})}. \tag{13}$$

Here, $\Lambda^{d-1} = (-\partial_t^2)^{\frac{d-1}{2}}$, $c_d^{-1} = 2(2\pi)^{d-1}$ is a dimension-dependent constant, and $\|\cdot\|_{\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})}$ denotes the total variation norm in the sense of measures supported on $\mathbb{S}^{d-1} \times \mathbb{R}$. Note that all operators must be understood in the distributional sense (see Parhi and Nowak (2021); Parhi and Unser (2024) for more details). The seminorm in Eq. (13) exactly coincides with the representational cost of a function realized as a single-hidden-layer bounded-norm infinite-width network and coincides with the $\mathcal{R}$-norm introduced by Ongie et al. (2020).

4. $\mathcal{R}\mathrm{TV}^2$ of a Kernel Machine

In this section, we study the $\mathcal{R}\mathrm{TV}^2$ of kernel machines in the RKHS $\mathcal{H}$. We show that one can write an explicit computable form for the case when the input dimension d is odd. Consider the underlying matrix $\mathbf{M} \succ 0$ for the Gaussian kernel $k_{\mathbf{M}}$ has the following Cholesky decomposition

$$\mathbf{M} = \mathbf{L}\mathbf{L}^{\mathsf{T}}. \tag{14}$$

Since $\mathbf{M}$ is full rank and is in $\mathrm{Sym}_+(\mathbb{R}^{d\times d})$ this decomposition is unique. With this we state the following result on $\mathcal{R}\mathrm{TV}^2(f)$ of a kernel machine $f \in \mathcal{H}(\mathbb{R}^d)$ with the proof in Appendix A.

Theorem 4 *Assume that the input dimension d is odd. For a kernel machine $f \in \mathcal{H}(\mathbb{R}^d)$ of the form*

$$f(\cdot) = \sum_{i=1}^{k} \alpha_i k_{\mathbf{M}}(\boldsymbol{x}_i, \cdot), \tag{15}$$

the $\mathcal{R}\mathrm{TV}^2$ of f is given by

$$\mathcal{R}\mathrm{TV}^2(f) = \frac{1}{|\det \mathbf{L}|}\frac{1}{\sqrt{2\pi}} \int_{\mathbb{S}^{d-1}} \frac{1}{\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\|} \int_{\mathbb{R}} \left| \sum_{i=1}^{k} \alpha_i \left(\frac{\partial^{d+1}}{\partial t^{d+1}} \exp\left(-\frac{(t - \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta})^2}{2\,\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\|^2} \right) \right) \right| \mathrm{d}t\, \mathrm{d}\boldsymbol{\beta}, \tag{16}$$

where we have used the decomposition $\mathbf{M} = \mathbf{L}^{\mathsf{T}}\mathbf{L}$. Furthermore, this can be extended to the case when f has a representation with infinite kernel functions by taking limits.

Proof Outline The proof proceeds in three main steps: First, we leverage the factorization $\mathbf{M} = \mathbf{L}^{\mathsf{T}}\mathbf{L}$ to express the Gaussian kernel for a single center $\boldsymbol{x}_0$ as

$$g(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}} \exp\Big(-\frac{|\mathbf{L}(\boldsymbol{x} - \boldsymbol{x}_0)|^2}{2} \Big). \tag{17}$$

Next, we compute its Fourier transform using the change-of-variables formula to obtain

$$\hat{g}(\boldsymbol{\omega}) = \exp\Big(-\mathrm{i}\boldsymbol{x}_0^{\mathsf{T}}\boldsymbol{\omega} \Big) \frac{1}{|\det \mathbf{L}|} \exp\Big(-\frac{|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\omega}|^2}{2} \Big). \tag{18}$$

Finally, we apply the Fourier slice theorem (Ramm and Katsevich, 1996) to connect the one-dimensional Fourier transform of $\mathcal{R}\{g\}(\boldsymbol{\beta}, t)$ (with respect to t) with d-variate Fourier transform evaluated on one slice: $\hat{g}(\omega\boldsymbol{\beta})$. By inverting this transform, we derive the explicit expression for $\mathcal{R}\{g\}(\boldsymbol{\beta}, t)$. For odd dimensions d, the second-order Radon total variation of smooth functions is characterized by the L_1-norm of the $(d+1)$th t-derivative of $\mathcal{R}\{g\}(\boldsymbol{\beta}, t)$ (cf., Ongie et al., 2020; Parhi and Nowak, 2021). The result then readily extends to any finite kernel machine

$$f(\cdot) = \sum_{i=1}^{k} \alpha_i k_{\mathbf{M}}(\boldsymbol{x}_i, \cdot) \tag{19}$$

through the linearity of both the Fourier and Radon transforms. ■

4.1. $\mathcal{R}\mathrm{TV}^2$ as an Expression of Hermite Polynomials

In Section 3, we discussed Hermite polynomials (probabilist's). In the following, we show how Theorem 4 can be rewritten in terms of Hermite polynomials. In the next section, we study certain useful property of this expression to show the construction of a diverging $\mathcal{R}\mathrm{TV}^2$ sequence of kernel machines.

First, consider the the case of a $g \in \mathcal{H}$ defined on one center $\boldsymbol{x}_0 \in \mathbb{R}^d$. Using Theorem 4 we can write the $\mathcal{R}\mathrm{TV}^2$-norm of the kernel machine g for one center $\boldsymbol{x}_0$ as

$$\mathcal{R}\mathrm{TV}^2(g) = \frac{1}{|\det \mathbf{L}|}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{S}^{d-1}} \frac{1}{\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\|}\int_{\mathbb{R}}\left|\left(\frac{\partial^{d+1}}{\partial t^{d+1}}\exp\left(-\frac{(t-\boldsymbol{x}_0^{\mathsf{T}}\boldsymbol{\beta})^2}{2\left\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\right\|^2}\right)\right)\right| \mathrm{d}t\,\mathrm{d}\boldsymbol{\beta}. \tag{20}$$

First, consider the inner integral in $\mathcal{R}\mathrm{TV}^2(g)$ and denote it as

$$I(\boldsymbol{\beta}) := \int_{\mathbb{R}}\left|\left(\frac{\partial^{d+1}}{\partial t^{d+1}}\exp\left(-\frac{(t-\boldsymbol{x}_0^{\mathsf{T}}\boldsymbol{\beta})^2}{2\sigma^2}\right)\right)\right| \mathrm{d}t, \tag{21}$$

where we use $\sigma = \left\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\right\|$.

Now, denote $\mu := \boldsymbol{x}_0\boldsymbol{\beta}$. Then, we note that the $(d+1)$-th derivative of $\exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)$ is related to the $(d+1)$-th Hermite polynomial H_{d+1} as follows:

$$\frac{\partial^{d+1}}{\partial t^{d+1}}\exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) = (-1)^{d+1}\sigma^{-(d+1)}H_{d+1}\left(\frac{t-\mu}{\sigma}\right)\exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right). \tag{22}$$

Now, let $u = \frac{t-\mu}{\sigma}$. Thus, $\mathrm{d}u = \frac{1}{\sigma}\mathrm{d}t$. Substituting this transformation to $I(\boldsymbol{\beta})$ gives

$$I(\boldsymbol{\beta}) = \int_{\mathbb{R}}\left|\left(\frac{\partial^{d+1}}{\partial t^{d+1}}\exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)\right)\right| \mathrm{d}t = \int_{\mathbb{R}}\left|(-1)^{d+1}\sigma^{-(d+1)}H_{d+1}(u)e^{-\frac{u^2}{2}}\right|\sigma\mathrm{d}u \tag{23}$$

$$= \sigma^{-d}\int_{\mathbb{R}}\left|H_{d+1}(u)e^{-\frac{u^2}{2}}\right|\mathrm{d}u. \tag{24}$$

We can rewrite $I(\boldsymbol{\beta})$ as

$$I(\boldsymbol{\beta}) = \sigma^{-d}\int_{\mathbb{R}}\left|H_{d+1}(u)e^{-\frac{u^2}{2}}\right|\mathrm{d}u = \sigma^{-d}C_d, \tag{25}$$

where $C_d := \int_{\mathbb{R}}\left|H_{d+1}(u)e^{-\frac{u^2}{2}}\right|\mathrm{d}u$. In Section 5, we bound this quantity to achieve certain decay of an infinite sum.

Replacing the computation in Eq. (24) to Eq. (20) gives

$$\mathcal{R}\mathrm{TV}^2(g) = \frac{C_d}{|\det \mathbf{L}|}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{S}^{d-1}}\frac{1}{\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\|^{d+1}}\mathrm{d}\boldsymbol{\beta}. \tag{26}$$

Thus, this shows that the expression of $\mathcal{R}\mathrm{TV}^2(g)$ in Theorem 4 can be simplified in terms of Hermite polynomials.

In the following we extend this for $k > 1$ with the proof deferred to Appendix B.

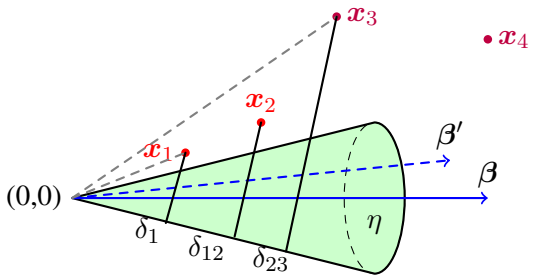

Figure 1: Illustration of an $(\boldsymbol{\beta}, \delta, \eta)$-separated set and a sequence $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4)$ that satisfy the requirements of the definition. The distances $\delta_1, \delta_{12}, \delta_{23}$ are at least δ apart.

Lemma 5 *For a kernel machine $f \in \mathcal{H}$ in the space of Gaussian RKHS. If f has the following representation*

$$f(\cdot) = \sum_{i=1}^{k} \alpha_i k_{\mathbf{M}}(\boldsymbol{x}_i, \cdot) \tag{27}$$

for a center set $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k\}$. Then, we can write

$$\mathcal{R}\mathrm{TV}^2(f) = \frac{1}{|\det \mathbf{L}|\sqrt{2\pi}} \int_{\mathbb{S}^{d-1}} \frac{I_{\mathit{inner}}(\boldsymbol{\beta})}{\sigma^{d+1}}\, d\boldsymbol{\beta}, \tag{28}$$

$$I_{\mathit{inner}}(\boldsymbol{\beta}) := \int_{\mathbb{R}} \left| \sum_{i=1}^{k} \alpha_i H_{d+1}\left(y + \Delta_i\right) e^{-\frac{(y+\Delta_i)^2}{2}} \right| dy, \tag{29}$$

where $\sigma = \|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\|$ and

$$\Delta_i = \frac{\boldsymbol{x}_1^{\mathsf{T}}\boldsymbol{\beta} - \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}{\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\|} \quad \textit{for } i = 2, 3, \ldots, k, \tag{30}$$

and $\Delta_1 = 0$.

5. A Sequence of Kernel Machines with Diverging $\mathcal{R}\mathrm{TV}^2$

In this section, we construct a sequence of kernel machines $\left\{f_n \in \mathcal{H}(\mathbb{R}^d)\right\}$ such that their $\mathcal{R}\mathrm{TV}^2$ diverges. First, we state some useful assumptions on the probabilist's Hermite polynomial which are easy to verify to hold in general (but surely in odd dimension d).

Assumption 1 (δ-peak) *Fix a dimension d. For a given Hermite polynomial H_{d+1}, we call an interval $[-\delta, \delta]$ a region of δ-peak if:*

1. $\frac{\partial H_{d+1}(y) e^{-\frac{y^2}{2}}}{\partial y} < 0$ *for all* $y > \delta$
2. $\frac{\partial H_{d+1}(y) e^{-\frac{y^2}{2}}}{\partial y} > 0$ *for all* $y < -\delta$

Due to exponential decay of the product $H_{d+1}(y)e^{-\frac{y^2}{2}}$, for any odd dimension d note that

$$\frac{\partial H_{d+1}(y)e^{-\frac{y^2}{2}}}{\partial y} = (H'_{d+1}(y) - yH_{d+1}(y))e^{-y^2/2}, \tag{31}$$

where $-yH_{d+1}(y)$ is a polynomial with odd dimension with negative highest term and this implies there exists a δ-peak. Now, we state a trivial observation on the absolute integral of $H_{d+1}(y)\ e^{-\frac{y^2}{2}}$.

Assumption 2 (ϵ-safe) *We say a constant $\epsilon > 0$ is ϵ-safe if*

$$\int_{[-\epsilon,\epsilon]} \left| H_{d+1}(y)\ e^{-\frac{y^2}{2}} \right| dy > 0. \tag{32}$$

Since H_{d+1} is non-zero polynomial this holds trivially for any $\epsilon > 0$. Furthermore, the integral is increasing with the size of an ϵ-interval. With this, we state a useful result on the convergence of a series of evaluations of $H_{d+1}(y)\ e^{-\frac{y^2}{2}}$ on distinct points $y \in \mathbb{R}$. The proof appears in Appendix C.

Lemma 6 *Let $d \geq 0$ be fixed and let $H_{d+1}(y)$ denote the Hermite polynomial of degree $d+1$. Then for any constant $\rho > 0$, there exists a constant $\delta_0 > 0$ (depending only on d) such that for every $\delta \geq \delta_0$ we have*

$$\sum_{j=2}^{\infty} \left| H_{d+1}(j\delta) \right| e^{-\frac{(j\delta)^2}{2}} < \frac{\rho}{4}. \tag{33}$$

5.1. Construction of a Diverging Sequence

In Section 3, we defined the notions of $(\boldsymbol{\beta}, \delta, \eta)$-separated sets of size $n \in \mathbb{N}$. Let $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ be a sequence in this set. Intuitively, any two centers in the sequence are at least δ apart when projected onto any direction $\boldsymbol{\beta}'$ such that $\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\beta}' \geq \|\boldsymbol{\beta}\|\,\|\boldsymbol{\beta}'\|\,\eta$ (see Fig. 1 for an illustration). Now, note that in Lemma 5, we provided an alternate representation of the $\mathcal{R}\mathrm{TV}^2(f)$ of a function as shown in Theorem 4, specifically the inner integral for each $\boldsymbol{\beta} \in \mathbb{S}^{d-1}$ has the form:

$$I_{\mathsf{inner}}(\boldsymbol{\beta}) := \int_{\mathbb{R}} \left| \sum_{i=1}^{k} \alpha_i H_{d+1}(y + \Delta_i)\, e^{-\frac{(y+\Delta_i)^2}{2}} \right| dy, \tag{34}$$

where each $\Delta_i = \boldsymbol{x}_1^\mathsf{T}\boldsymbol{\beta} - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}$ (ignoring the normalization). If the projections $\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}$ are far apart on the real line $\mathbb{R}$, noting the absolute decay in the values of $H_{d+1}(y)\ e^{-\frac{y^2}{2}}$ outside the region of δ-peak as asserted by Assumption 1, we can quantify and control contributions of terms corresponding to $j \neq i$ in the inner integral.

Now, the property holds over a non-trivial cone $\mathcal{K}(\boldsymbol{\beta}) := \left\{\boldsymbol{\beta}' \in \mathbb{S}^{d-1} \,|\, \boldsymbol{\beta}^\mathsf{T}\boldsymbol{\beta}' \geq \eta\right\}$ with non-zero volume. Now, note in Eq. (28), which involves the following integral

$$\int_{\mathbb{S}^{d-1}} \frac{I_{\mathsf{inner}}(\boldsymbol{\beta})}{\sigma^{d+1}}\, d\boldsymbol{\beta} \tag{35}$$

is non-trivially positive. Thus, we show along any direction $\boldsymbol{\beta}$ in the cone, I_{inner} diverges as k grows if the kernel machine f is defined for a sequence of centers from the $(\boldsymbol{\beta}, \delta, \eta)$-separated set.

Now, we state the main theorem of the work. For ease of analysis we assume that the largest eigenvalue of $\mathbf{L}^{-1}$ is upper bounded by 1 which can be easily replaced with appropriate rescaling and choice of the parameters in the statement.

Theorem 7 (Diverging $\mathcal{R}\mathrm{TV}^2$) *Consider the Gaussian RKHS $\mathcal{H}(\mathbb{R}^d)$ as defined in Eq. (4). Assume $\epsilon \in (0, 1/2]$ be a safe constant (see Assumption 2). Define*

$$\rho := \int_{[-\epsilon,\epsilon]} \left| H_{d+1}(y)\, e^{-\frac{y^2}{2}} \right| dy. \tag{36}$$

Fix a unit vector $\boldsymbol{\beta} \in \mathbb{R}^d$, scalars $\eta \geq \frac{\sqrt{3}}{2}$, and $\delta = 3\max\{\epsilon, \delta_0(\rho), \delta'\}$ where $\delta_0(\rho)$ is chosen as per Lemma 6, and $\delta'(d)$ as per Assumption 1. Let $\mathcal{X}_\infty = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\} \subset \mathbb{R}^d$ be an infinite sequence such that any subsequence $\Gamma_n = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ is in the $(\boldsymbol{\beta}, \delta, \eta)$-separated set of size n. Define a function $f \in \mathcal{H}$ on $\mathcal{X}_\infty$ that has a representation with an f-unbounded combination α_f. Then,

$$\mathcal{R}\mathrm{TV}^2(\{f_n\}) \to \infty, \tag{37}$$

as $n \to \infty$.

Proof First, we rewrite Eq. (28) for the $\mathcal{R}\mathrm{TV}^2$ of the function f_k as follows

$$\mathcal{R}\mathrm{TV}^2(f_k) = \frac{1}{|\det \mathbf{L}|\sqrt{2\pi}} \int_{\mathbb{S}^{d-1}} \frac{1}{\sigma^{d+1}} \int_{\mathbb{R}} \left| \sum_{i=1}^{k} \alpha_i H_{d+1}(y + \Delta_i)\, e^{-\frac{(y+\Delta_i)^2}{2}} \right| dy\, d\boldsymbol{\beta}, \tag{38}$$

with the inner integral

$$I_{\mathsf{inner}}(\boldsymbol{\beta}) = \int_{\mathbb{R}} \left| \sum_{i=1}^{k} \alpha_i\, H_{d+1}(y + \Delta_i)\, e^{-\frac{(y+\Delta_i)^2}{2}} \right| dy, \tag{39}$$

where

$$\Delta_1 = 0,\ \Delta_i = \frac{a_1 - a_i}{\sigma} = \frac{\boldsymbol{x}_1^\mathsf{T}\boldsymbol{\beta} - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}}{\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\|} \quad \text{for } i = 2, 3, \ldots, k. \tag{40}$$

First, define the cone $\mathcal{K}$ wrt $\boldsymbol{\beta}$ and η as stated in the theorem statement, i.e.,

$$\mathcal{K} := \left\{ \boldsymbol{\beta}' \in \mathbb{S}^{d-1} \,\middle|\, \boldsymbol{\beta}'^\mathsf{T}\boldsymbol{\beta} \geq \eta \right\}. \tag{41}$$

Note that the volume $\mathsf{vol}(\mathcal{K}) > 0$ implying that

$$\int_{\mathbb{S}^{d-1}} \frac{1}{\sigma^{d+1}}\, d\boldsymbol{\beta} \geq \int_{\mathcal{K}} \frac{1}{\sigma^{d+1}}\, d\boldsymbol{\beta} = (\star). \tag{42}$$

Note that $\mathbf{M} = \mathbf{L}^\mathsf{T}\mathbf{L}$. We assume that the $\mathbf{M}$ is symmetric and PSD, implying that singular values of $\mathbf{L}$ are *exactly* the square root of the eigenvalues of $\mathbf{M}$, i.e.,

$$\sigma_i(\mathbf{L}) = \sqrt{|\lambda_i(\mathbf{M})|}. \tag{43}$$

But, since $\mathbf{L}$ is invertible implying the singular values if $\mathbf{L}^{-1}$ are inverses to singular values of $\mathbf{L}$, i.e., $\sigma_i(\mathbf{L}^{-1}) = \frac{1}{\sigma_i(\mathbf{L})}$. Thus, we can rewrite Eq. (42) as

$$\begin{aligned}(\star) &= \int_{\mathcal{K}} \frac{1}{\sigma_{\max}(\mathbf{L}^{-1})^{d+1}}\, d\boldsymbol{\beta} \\ &\geq \int_{\mathcal{K}} \sigma_{\min}(\mathbf{L})^{d+1}\, d\boldsymbol{\beta} \\ &= \int_{\mathcal{K}} \lambda_{\min}(\mathbf{M})^{\frac{(d+1)}{2}}\, d\boldsymbol{\beta} = \lambda_{\min}(\mathbf{M})^{\frac{(d+1)}{2}} \mathsf{vol}(\mathcal{K}) > 0. \end{aligned} \tag{44}$$

Now, we will show for any $\boldsymbol{\beta}' \in \mathcal{K}$, there is a non-trivial lower bound on $I_{\mathsf{inner}}(\boldsymbol{\beta}')$. Note that by definition, Γ_n is in the $(\boldsymbol{\beta}', \delta)$-separated set.

Hence, for all $i, j = 2, 3, \ldots$

$$|\Delta_i - \Delta_j| \geq \delta. \tag{45}$$

Define the neighborhoods $\{N_i\}$ for the safe constant ϵ as follows

$$N_i := [-\Delta_i - \epsilon, -\Delta_i + \epsilon]\,. \tag{46}$$

Now, consider the integral on the neighborhood N_i:

$$\begin{aligned}&\int_{N_i} \left| \sum_{i=1}^{k} \alpha_i\, H_{d+1}\left(y + \Delta_i\right)\, e^{-\frac{(y+\Delta_i)^2}{2}} \right| dy \\ &\geq \int_{N_i} \left| \alpha_i\, H_{d+1}\left(y + \Delta_i\right)\, e^{-\frac{(y+\Delta_i)^2}{2}} \right| dy - \underbrace{\int_{N_i} \left| \sum_{j=1, j\neq i}^{k} \alpha_j\, H_{d+1}\left(y + \Delta_j\right)\, e^{-\frac{(y+\Delta_j)^2}{2}} \right| dy}_{=:\theta_i}. \end{aligned} \tag{47}$$

The second line follows from the triangle inequality. Now, change of variable simplifies the first equation as

$$\int_{N_i} \left| \alpha_i\, H_{d+1}\left(y + \Delta_i\right)\, e^{-\frac{(y+\Delta_i)^2}{2}} \right| dy \geq |\alpha_i| \int_{[-\epsilon,\epsilon]} \left| H_{d+1}\left(y\right)\, e^{-\frac{y^2}{2}} \right| dy \geq |\alpha_i|\rho. \tag{48}$$

In the last equation, we used the definition of ρ.

Now, summing over each $i = 1, 2, \ldots k$, we get

$$I_{\mathsf{inner}} \geq \sum_{i=1}^{k} \int_{N_i} \left| \sum_{i=1}^{k} \alpha_i\, H_{d+1}\left(y + \Delta_i\right)\, e^{-\frac{(y+\Delta_i)^2}{2}} \right| dy \geq \rho \sum_{i=1}^{k} |\alpha_i| - \sum_{i=1}^{k} \theta_i. \tag{49}$$

Now, we will show how to bound the sum $\sum_{i=1}^{k} \theta_i$.

Bounding θ_k: First note that, we can bound each θ_i as follows

$$\begin{aligned}
\theta_i &= \int_{N_i} \left| \sum_{j=1, j\neq i}^{k} \alpha_j \, H_{d+1}\left(y+\Delta_j\right) \, e^{-\frac{(y+\Delta_j)^2}{2}} \right| dy \\
&\leq \sum_{j=1, j\neq i}^{k} \int_{N_i} \left| \alpha_j \, H_{d+1}\left(y+\Delta_j\right) \, e^{-\frac{(y+\Delta_j)^2}{2}} \right| dy \qquad (50) \\
&\leq \sum_{j=1, j\neq i}^{k} \int_{[-\epsilon,\epsilon]} \left| \alpha_j \, H_{d+1}\left(-\Delta_i+y+\Delta_j\right) \, e^{-\frac{(-\Delta_i+y+\Delta_j)^2}{2}} \right| dy \qquad (51) \\
&\leq \sum_{j=1, j\neq i}^{k} |\alpha_j| \int_{[-\epsilon,\epsilon]} \left| H_{d+1}\left(|i-j|\delta\right) \, e^{-\frac{(|i-j|\delta)^2}{2}} \right| dy \qquad (52) \\
&\leq 2\epsilon \sum_{j=1, j\neq i}^{k} |\alpha_j| \left| H_{d+1}\left(|i-j|\delta\right) \right| \, e^{-\frac{(|i-j|\delta)^2}{2}} \\
&\leq \sum_{j=1, j\neq i}^{k} |\alpha_j| \left| H_{d+1}\left(|i-j|\delta\right) \right| \, e^{-\frac{(|i-j|\delta)^2}{2}}. \qquad (53)
\end{aligned}$$

Eq. (50) is a straight-forward application of triangle inequality. In Eq. (51), we simplify Eq. (50) via change of variable. In Eq. (52), we use the assumption of δ-peak. To simplify $-\Delta_i + y + \Delta_j$ for a choice of $y \in [-\epsilon, \epsilon]$, we assume that indices of the projections Δ_i for $i = 1, 2, \ldots$ are arranged in ascending order in their values on the real line. Since each consecutive projections are at least δ apart, we can bound $|-\Delta_i + y + \Delta_j| > (|i-j| - 1/3)\delta$. Since Hermite polynomials in even dimension, i.e. $d+1$, are even

$$\left| H_{d+1}\left(-\Delta_i+y+\Delta_j\right) \, e^{-\frac{(-\Delta_i+y+\Delta_j)^2}{2}} \right| \leq \left| H_{d+1}\left((|i-j|-2/3)\delta\right) e^{-\frac{((|i-j|-2/3)\delta)^2}{2}} \right|. \qquad (54)$$

For simplification, we have omitted the $-(2/3)\delta$ additive term in the equation above. Finally, Eq. (53) follows as $\epsilon \leq 1/2$.

Summing over each $i = 1, \ldots, k$

$$\begin{aligned}
\sum_{i=1}^{k} \theta_k &= \sum_{i=1}^{k} \int_{N_i} \left| \sum_{j=1, j\neq i}^{k} \alpha_j \, H_{d+1}\left(y+\Delta_j\right) \, e^{-\frac{(y+\Delta_j)^2}{2}} \right| dy \\
&\leq \sum_{i=1}^{k} \sum_{j=1, j\neq i}^{k} |\alpha_j| \left| H_{d+1}\left(|i-j|\delta\right) \right| \, e^{-\frac{(|i-j|\delta)^2}{2}} \\
&\leq 2 \left(\sum_{j=1}^{k} \left| H_{d+1}\left(j\delta\right) \right| \, e^{-\frac{(j\delta)^2}{2}} \right) \sum_{i=1}^{k} |\alpha_i|. \qquad (55)
\end{aligned}$$

Using Lemma 6, we can rewrite Eq. (49) as

$$I_{\text{inner}} \geq \rho \sum_{i=1}^{k} |\alpha_i| - 2 \left(\sum_{j=2}^{k} \left| H_{d+1}\left(j\delta\right) \right| \, e^{-\frac{(j\delta)^2}{2}} \right) \sum_{i=1}^{k} |\alpha_i| \geq \frac{\rho}{2} \sum_{i=1}^{k} |\alpha_i|. \qquad (56)$$

Now, note that using Eq. (42) and Eq. (44)

$$\begin{aligned}\mathcal{R}\mathrm{TV}^2(f_k) &\geq \frac{1}{|\det \mathbf{L}|\sqrt{2\pi}} \int_{\mathcal{K}} \frac{1}{\sigma^{d+1}} \int_{\mathbb{R}} \left| \sum_{i=1}^{k} \alpha_i H_{d+1}\left(y+\Delta_i\right) e^{-\frac{(y+\Delta_i)^2}{2}} \right| dy\, d\boldsymbol{\beta} \\ &\geq \frac{1}{|\det \mathbf{L}|\sqrt{2\pi}} \int_{\mathcal{K}} \frac{1}{\sigma^{d+1}} \left(\frac{\rho}{2} \sum_{i=1}^{k} |\alpha_i| \right) d\boldsymbol{\beta} \\ &\geq \left(\frac{1}{|\det \mathbf{L}|\sqrt{2\pi}} \lambda_{\min}(\mathbf{M})^{\frac{(d+1)}{2}} \mathsf{vol}(\mathcal{K}) \right) \cdot \left(\frac{\rho}{2} \sum_{i=1}^{k} |\alpha_i| \right). \end{aligned} \tag{57}$$

Now, in the limiting case

$$\begin{aligned}\lim_{k\to\infty} \mathcal{R}\mathrm{TV}^2(f_k) &\geq \left(\frac{1}{|\det \mathbf{L}|\sqrt{2\pi}} \lambda_{\min}(\mathbf{M})^{\frac{(d+1)}{2}} \mathsf{vol}(\mathcal{K}) \right) \lim_{k\to\infty} \left(\frac{\rho}{2} \sum_{i=1}^{k} |\alpha_i| \right) \\ &= \left(\frac{1}{|\det \mathbf{L}|\sqrt{2\pi}} \lambda_{\min}(\mathbf{M})^{\frac{(d+1)}{2}} \mathsf{vol}(\mathcal{K}) \right) \cdot \frac{\rho}{2} \|\alpha_f\| \to \infty. \end{aligned} \tag{58}$$

Hence the claim of the theorem has been proven. ■

6. Conclusion

In this work, we showed that if we allow *unbounded* domains, there exist functions that cannot be represented within $\mathcal{R}\mathrm{BV}^2(\mathbb{R}^d)$, but can be represented within the Gaussian RKHS. This analysis reveals a nontrivial gap between kernel methods and neural networks by exhibiting functions that kernel methods can represent, whereas neural networks cannot. On that note, this observation motivates further investigation of what this gap entails in a learning setting. This observation also motivates investigating if a similar gap exists between kernel methods and *deep* neural networks. We leave the details to future work.

Acknowledgement

Authors thank anonymous reviewers for helpful feedback on the work. AK thanks the National Science Foundation for support under grant IIS-2211386 in the duration of this project. MB acknowledges support from the National Science Foundation (NSF) and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning through awards DMS-2031883 and #814639 as well as the TILOS institute (NSF CCF-2112665) and the Office of Naval Research (ONR N000142412631).

References

Zeyuan Allen-Zhu and Yuanzhi Li. What can resnet learn efficiently, going beyond kernels? *Advances in Neural Information Processing Systems*, 32, 2019.

Sanjeev Arora, Simon S Du, Zhiyuan Li, Ruslan Salakhutdinov, Ruosong Wang, and Dingli Yu. Harnessing the power of infinitely wide deep nets on small-data tasks. *arXiv preprint arXiv:1910.01663*, 2019.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *J. Mach. Learn. Res.*, 18(1):629–681, January 2017. ISSN 1532-4435.

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

Francesca Bartolucci, Ernesto De Vito, Lorenzo Rosasco, and Stefano Vigogna. Understanding neural networks with reproducing kernel banach spaces. *Applied and Computational Harmonic Analysis*, 62:194–236, 2023. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2022.08.006. URL https://www.sciencedirect.com/science/article/pii/S1063520322000768.

Alberto Bietti and Francis Bach. Deep equals shallow for ReLU networks in kernel regimes. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=aDjoksTpXOP.

Lin Chen and Sheng Xu. Deep neural tangent kernel and laplace kernel have the same {rkhs}. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=vK9WrZ0QYQ.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural network. *Advances in Neural Information Processing Systems*, 32, 2019.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods?*. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124009, dec 2021a. doi: 10.1088/1742-5468/ac3a81. URL https://dx.doi.org/10.1088/1742-5468/ac3a81.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029 – 1054, 2021b. doi: 10.1214/20-AOS1990. URL https://doi.org/10.1214/20-AOS1990.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences, 2018. URL https://arxiv.org/abs/1807.02582.

Tong Mao, Jonathan W. Siegel, and Jinchao Xu. Approximation rates for shallow relu^k neural networks on sobolev spaces via the radon transform, 2024. URL https://arxiv.org/abs/2408.10996.

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for non-convex losses. *Annals of Statistics*, 46, 07 2016. doi: 10.1214/17-AOS1637.

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. doi: 10.1073/pnas.1806579115. URL https://www.pnas.org/doi/abs/10.1073/pnas.1806579115.

Roman Novak, Lechao Xiao, Jaehoon Lee, Yasaman Bahri, Greg Yang, Jiri Hron, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Bayesian deep convolutional networks with many channels are gaussian processes. *arXiv preprint arXiv:1810.05148*, 2018.

Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width relu nets: The multivariate case. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1lNPxHKDH.

Rahul Parhi and Robert D. Nowak. Banach space representer theorems for neural networks and ridge splines. *Journal of Machine Learning Research*, 22(43):1–40, 2021. URL https://jmlr.org/papers/v22/20-583.html.

Rahul Parhi and Robert D. Nowak. What kinds of functions do deep neural networks learn? insights from variational spline theory. *SIAM Journal on Mathematics of Data Science*, 4(2):464–489, 2022. doi: 10.1137/21M1418642.

Rahul Parhi and Robert D. Nowak. Near-minimax optimal estimation with shallow ReLU neural networks. *IEEE Transactions on Information Theory*, 69(2):1125–1140, 2023a. doi: 10.1109/TIT.2022.3208653.

Rahul Parhi and Robert D. Nowak. Deep learning meets sparse regularization: A signal processing perspective. *IEEE Signal Processing Magazine*, 40(6):63–74, 2023b. doi: 10.1109/MSP.2023.3286988.

Rahul Parhi and Michael Unser. Distributional extension and invertibility of the k-plane transform and its dual. *SIAM Journal on Mathematical Analysis*, 56(4):4662–4686, 2024. doi: 10.1137/23M1556721.

Rahul Parhi and Michael Unser. Function-space optimality of neural architectures with multivariate nonlinearities. *SIAM Journal on Mathematics of Data Science*, 7(1):110–135, 2025. doi: 10.1137/23M1620971.

Leonardo Petrini, Francesco Cagnetta, Eric Vanden-Eijnden, and Matthieu Wyart. Learning sparse features can lead to overfitting in neural networks*. *Journal of Statistical Mechanics: Theory and Experiment*, 2023(11):114003, nov 2023. doi: 10.1088/1742-5468/ad01b9. URL https://dx.doi.org/10.1088/1742-5468/ad01b9.

Alexander G. Ramm and Alexander I. Katsevich. *The Radon transform and local tomography*. CRC Press, Boca Raton, FL, 1996. ISBN 0-8493-9492-9.

Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75 (9):1889–1935, 2022.

Vaishaal Shankar, Alex Fang, Wenshuo Guo, Sara Fridovich-Keil, Jonathan Ragan-Kelley, Ludwig Schmidt, and Benjamin Recht. Neural kernels without tangents. In *International conference on machine learning*, pages 8614–8623. PMLR, 2020.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387772413.

Ingo Steinwart, Don R. Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *Annual Conference Computational Learning Theory*, 2009. URL https://api.semanticscholar.org/CorpusID:7741716.

G. Szegő. *Orthogonal Polynomials*. American Math. Soc: Colloquium publ. American Mathematical Society, 1975. ISBN 9780821810231. URL https://books.google.com/books?id=ZOhmnsXlcY0C.

Ulrike von Luxburg and Olivier Bousquet. Distance–based classification with lipschitz functions. *J. Mach. Learn. Res.*, 5:669–695, December 2004. ISSN 1532-4435.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. *Advances in neural information processing systems*, 32, 2019.

Ding-Xuan Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, 49(7):1743–1752, 2003. doi: 10.1109/TIT.2003.813564.

A Gap Between the Gaussian RKHS and Neural Networks: Supplementary Materials

Appendix A. $\mathcal{R}\mathrm{TV}^2$ for centers size $k > 1$

In this Appendix, we provide the proof of Theorem 4. First, we provide the proof for one center and then extend it to multi-center settings.

Proof Let

$$g(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_0\|_{\mathbf{M}}^2}{2}\right). \tag{59}$$

Also, define the $\mathbf{0}$ mean identity covariance Gaussian

$$g_0(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\|\boldsymbol{x}\|^2}{2}\right). \tag{60}$$

If we can write $\mathbf{M} = \mathbf{L}^\mathsf{T}\mathbf{L}$, then we have that

$$\|\boldsymbol{x} - \boldsymbol{x}_0\|_{\mathbf{M}}^2 = \|\mathbf{L}(\boldsymbol{x} - \boldsymbol{x}_0)\|^2, \tag{61}$$

in which case

$$g(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\|\mathbf{L}\boldsymbol{x} - \mathbf{L}\boldsymbol{x}_0\|^2}{2}\right). \tag{62}$$

We have the Fourier transform

$$\hat{g}_0(\boldsymbol{\omega}) = \exp\left(-\frac{\|\boldsymbol{\omega}\|^2}{2}\right). \tag{63}$$

We have the equality $g(\boldsymbol{x}) = g_0(\mathbf{L}\boldsymbol{x} - \mathbf{L}\boldsymbol{x}_0)$. Using the change of variables formula for the Fourier transform, we have

$$\hat{g}(\boldsymbol{\omega}) = \exp\left(-\mathrm{i}(\mathbf{L}\boldsymbol{x}_0)^\mathsf{T}\mathbf{L}^{-\mathsf{T}}\boldsymbol{\omega}\right) \frac{1}{|\det \mathbf{L}|}\hat{g}_0(\mathbf{L}^{-\mathsf{T}}\boldsymbol{\omega}) \tag{64}$$

$$= \exp\left(-\mathrm{i}\boldsymbol{x}_0^\mathsf{T}\mathbf{L}^\mathsf{T}\mathbf{L}^{-\mathsf{T}}\boldsymbol{\omega}\right) \frac{1}{|\det \mathbf{L}|} \exp\left(-\frac{\left\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\omega}\right\|^2}{2}\right) \tag{65}$$

$$= \exp\left(-\mathrm{i}\boldsymbol{x}_0^\mathsf{T}\boldsymbol{\omega}\right) \frac{1}{|\det \mathbf{L}|} \exp\left(-\frac{\left\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\omega}\right\|^2}{2}\right). \tag{66}$$

The Fourier slice theorem (Ramm and Katsevich, 1996) says that

$$\mathcal{F}_1\{\mathcal{R}\{f\}(\boldsymbol{\beta}, \cdot)\}(\omega) = \hat{f}(\omega\boldsymbol{\beta}). \tag{67}$$

If we evaluate $\hat{g}$ at $\boldsymbol{\omega} = \omega\boldsymbol{\beta}$ in the Eq. (66), we find

$$\hat{g}(\omega\boldsymbol{\beta}) = \exp\left(-\mathrm{i}\boldsymbol{x}_0^\mathsf{T}(\omega\boldsymbol{\beta})\right) \frac{1}{|\det \mathbf{L}|} \exp\left(-\frac{\left\|\mathbf{L}^{-\mathsf{T}}(\omega\boldsymbol{\beta})\right\|^2}{2}\right) \tag{68}$$

$$= \exp\left(-\mathrm{i}(\boldsymbol{x}_0^\mathsf{T}\boldsymbol{\beta})\omega\right) \frac{1}{|\det \mathbf{L}|} \exp\left(-\frac{|\omega|^2\left\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\right\|^2}{2}\right). \tag{69}$$

The 1D inverse Fourier transform of this is the Radon transform of g, i.e.,

$$\mathcal{R}\{g\}(\boldsymbol{\beta}, t) = \frac{1}{|\det \mathbf{L}|} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\|^2}} \exp\left(-\frac{(t - \boldsymbol{x}_0^{\mathsf{T}}\boldsymbol{\beta})^2}{2\,\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\|^2}\right). \tag{70}$$

If d is odd, then the second-order Radon domain total variation is the L_1-norm of $(d+1)$ derivatives in t of this quantity (see Equation (28) in Parhi and Nowak (2021)). That is

$$\mathcal{R}\mathrm{TV}^2(g) = \frac{1}{|\det \mathbf{L}|} \frac{1}{\sqrt{2\pi}} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \left| \frac{1}{\sqrt{\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\|^2}} \left(\frac{\partial^{d+1}}{\partial t^{d+1}} \exp\left(-\frac{(t - \boldsymbol{x}_0^{\mathsf{T}}\boldsymbol{\beta})^2}{2\,\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\|^2}\right)\right)\right| \mathrm{d}t\,\mathrm{d}\boldsymbol{\beta} \tag{71}$$

$$= \frac{1}{|\det \mathbf{L}|} \frac{1}{\sqrt{2\pi}} \int_{\mathbb{S}^{d-1}} \frac{1}{\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\|} \int_{\mathbb{R}} \left| \left(\frac{\partial^{d+1}}{\partial t^{d+1}} \exp\left(-\frac{(t - \boldsymbol{x}_0^{\mathsf{T}}\boldsymbol{\beta})^2}{2\,\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\|^2}\right)\right)\right| \mathrm{d}t\,\mathrm{d}\boldsymbol{\beta}. \tag{72}$$

This gives the stated expression on $\mathcal{R}\mathrm{TV}^2(f)$ for one center. ■

A.1. Multi-Center Computation

For a kernel machine with $k > 1$ centers, we can rewrite g as

$$g(\boldsymbol{x}) = \sum_{i=1}^{k} \frac{1}{(2\pi)^{d/2}} \alpha_i \cdot \exp\left(-\frac{\|\mathbf{L}\boldsymbol{x} - \mathbf{L}\boldsymbol{x}_i\|^2}{2}\right). \tag{73}$$

Denote by $g_i(\boldsymbol{x}) := \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\|\mathbf{L}\boldsymbol{x} - \mathbf{L}\boldsymbol{x}_i\|^2}{2}\right)$ for each center $\boldsymbol{x}_i \in \mathcal{D}$.

Now, the Fourier transform of g can be written for the extended case, noting the linearity of the transform,

$$\hat{g}(\boldsymbol{\omega}) = \sum_{i=1}^{k} \hat{g}_i(\boldsymbol{\omega}). \tag{74}$$

This implies that

$$\hat{g}(\boldsymbol{\omega}) = \sum_{i=1}^{k} \exp\left(-\mathrm{i}\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\omega}\right) \frac{1}{|\det \mathbf{L}|} \exp\left(-\frac{\left\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\omega}\right\|^2}{2}\right). \tag{75}$$

Now, computing the inverse Fourier transform of $\hat{g}$ wrt $\boldsymbol{\omega}$ gives

$$\mathcal{R}\{g\}(\boldsymbol{\beta}, t) = \frac{1}{|\det \mathbf{L}|} \sum_{i=1}^{k} \alpha_i \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\|^2}} \exp\left(-\frac{(t - \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta})^2}{2\,\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\|^2}\right). \tag{76}$$

As before the $\mathcal{R}\mathrm{TV}^2$ of g, i.e. the second-order Radon domain total variation for odd values of d is the L_1-norm of $(d+1)$ derivatives in t of this quantity. Thus,

$$\mathcal{R}\mathrm{TV}^2(g) = \frac{1}{|\det \mathbf{L}|} \frac{1}{\sqrt{2\pi}} \int_{\mathbb{S}^{d-1}} \frac{1}{\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\|} \int_{\mathbb{R}} \left| \sum_{i=1}^{k} \alpha_i \left(\frac{\partial^{d+1}}{\partial t^{d+1}} \exp\left(-\frac{(t - \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta})^2}{2\,\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\|^2}\right)\right)\right| \mathrm{d}t\,\mathrm{d}\boldsymbol{\beta}. \tag{77}$$

This expression can be extended to the case of kernel machines with infinite centers by taking limits. In particular, it is guaranteed to be finite for the case when the ℓ_1 norm of the coefficients α is finite since

$$\left|\sum_i \alpha_i \left(\frac{\partial^{d+1}}{\partial t^{d+1}} \exp\left(-\frac{(t-\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta})^2}{2\left\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\right\|^2}\right)\right)\right| \leq C \cdot \sum_i |\alpha_i|, \tag{78}$$

where it is straightforward to show that $\frac{\partial^{d+1}}{\partial t^{d+1}} \exp\left(-\frac{(t-\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta})^2}{2\left\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\right\|^2}\right)$ is bounded by a universal constant $C > 0$ for all choices of $\boldsymbol{x}_i$.

Appendix B. Change of Variables For Multiple Centers

In this Appendix, we provide the proof of Lemma 5.

Proof Previously, we computed the $\mathcal{R}\mathrm{TV}^2$ of a general kernel machine as

$$\mathcal{R}\mathrm{TV}^2(g) = \frac{1}{|\det \mathbf{L}|}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{S}^{d-1}} \frac{1}{\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\|}\int_{\mathbb{R}}\left|\sum_{i=1}^k \alpha_i \left(\frac{\partial^{d+1}}{\partial t^{d+1}} \exp\left(-\frac{(t-\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta})^2}{2\left\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\right\|^2}\right)\right)\right| \mathrm{d}t\,\mathrm{d}\boldsymbol{\beta}. \tag{79}$$

Now, we can rewrite the $(d+1)$-th derivative of the involved exponential as follows

$$\frac{\partial^{d+1}}{\partial t^{d+1}} \exp\left(-\frac{(t-a_i)^2}{2\sigma^2}\right) = (-1)^{d+1}\sigma^{-(d+1)} H_{d+1}\left(\frac{t-a_i}{\sigma}\right)\exp\left(-\frac{(t-a_i)^2}{2\sigma^2}\right),$$

where we define

$$a_i = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} \quad \text{and} \quad \sigma = \|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\|. \tag{80}$$

Substituting the expression for Hermite polynomial (see Section 3) into the integral I gives

$$I = \frac{1}{|\det \mathbf{L}|\sqrt{2\pi}}\int_{\mathbb{S}^{d-1}} \frac{1}{\sigma^{d+2}}\int_{\mathbb{R}}\left|\sum_{i=1}^k \alpha_i H_{d+1}\left(\frac{t-a_i}{\sigma}\right) e^{-\frac{(t-a_i)^2}{2\sigma^2}}\right| dt\, d\boldsymbol{\beta}.$$

To simplify the inner integral, we can perform the following change of variable centered at $a_1 = \boldsymbol{x}_1^\mathsf{T}\boldsymbol{\beta}$

$$y = \frac{t-a_1}{\sigma} \quad \Rightarrow \quad t = \sigma y + a_1 \quad \Rightarrow \quad dt = \sigma\, dy. \tag{81}$$

We can express all y_i in terms of y as follows

$$y_i = \frac{t-a_i}{\sigma} = \frac{\sigma y + a_1 - a_i}{\sigma} = y + \Delta_i, \tag{82}$$

where we define

$$\Delta_i = \frac{a_1 - a_i}{\sigma} = \frac{\boldsymbol{x}_1^\mathsf{T}\boldsymbol{\beta} - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}}{\|\mathbf{L}^{-\mathsf{T}}\boldsymbol{\beta}\|} \quad \text{for } i = 2, 3, \ldots, k, \tag{83}$$

and $\Delta_1 = 0$.

With this change of variable into the inner integral we have the stated final form

$$I = \frac{1}{|\det \mathbf{L}|\sqrt{2\pi}} \int_{\mathbb{S}^{d-1}} \frac{1}{\sigma^{d+1}} \underbrace{\int_{\mathbb{R}} \left| \sum_{i=1}^{k} \alpha_i H_{d+1}\left(y+\Delta_i\right) e^{-\frac{(y+\Delta_i)^2}{2}} \right| dy}_{\text{define } I_{\text{inner}}} \, d\boldsymbol{\beta} \tag{84}$$

$$I_{\text{inner}} = \int_{\mathbb{R}} \left| \sum_{i=1}^{k} \alpha_i \, H_{d+1}\left(y+\Delta_i\right) \, e^{-\frac{(y+\Delta_i)^2}{2}} \right| dy. \tag{85}$$

■

Appendix C. A Useful Property of Hermite Polynomials

In this Appendix, we provide the proof of Lemma 6.

Proof Since $H_{d+1}(y)$ is a polynomial of degree $d+1$, there exists a constant $C > 0$ (depending only on d) such that

$$\left| H_{d+1}(y) \right| \leq C \left(1 + |y|\right)^{d+1} \quad \text{for all } y \in \mathbb{R}. \tag{86}$$

Hence, for any $\delta > 0$ and any integer $j \geq 2$ we have

$$\left| H_{d+1}\big(j\delta\big) \right| e^{-\frac{(j\delta)^2}{2}} \leq C \left(1 + j\delta\right)^{d+1} e^{-\frac{(j\delta)^2}{2}}. \tag{87}$$

Now, we define

$$S(\delta) := \sum_{j=2}^{\infty} (1 + j\delta)^{d+1} \, e^{-\frac{(j\delta)^2}{2}}. \tag{88}$$

Note that we we can upper bound as follows

$$\sum_{j=2}^{\infty} \left| H_{d+1}\big(j\delta\big) \right| e^{-\frac{(j\delta)^2}{2}} \leq C \, S(\delta). \tag{89}$$

For each fixed $j \geq 2$, notice that $(1+j\delta)^{d+1} \, e^{-\frac{(j\delta)^2}{2}}$ decays exponentially in j (since the exponential term $e^{-\frac{(j\delta)^2}{2}}$ dominates the polynomial growth of $(1 + j\delta)^{d+1}$). Moreover, for fixed $j \geq 2$ we have

$$\lim_{\delta \to \infty} (1 + j\delta)^{d+1} \, e^{-\frac{(j\delta)^2}{2}} = 0. \tag{90}$$

Thus, the series $S(\delta)$ converges for every fixed $\delta > 0$ and

$$\lim_{\delta \to \infty} S(\delta) = 0. \tag{91}$$

Hence, by the definition, there exists some $\delta_0 > 0$ such that for all $\delta \geq \delta_0$ we have

$$S(\delta) < \frac{\rho}{4C}. \tag{92}$$

It follows that for every $\delta \geq \delta_0$,

$$\sum_{j=2}^{\infty} \left| H_{d+1}\big(j\delta\big) \right| e^{-\frac{(j\delta)^2}{2}} \leq C \cdot S(\delta) < C \cdot \frac{\rho}{4C} = \frac{\rho}{4}. \tag{93}$$

Thus for this choice of δ_0 we achieve the statement of the lemma. ■

Appendix D. A Sequence With Diverging ℓ_1-Norm and Converging RKHS Norm

In this Appendix, we provide the proof of Example 2.

Lemma 8 *Let $\alpha_n = \frac{1}{n}$ and suppose that the points $\boldsymbol{x}_n \in \mathbb{R}^d$ satisfy*

$$\|\boldsymbol{x}_i - \boldsymbol{x}_j\| \geq |i-j|\delta \quad \textit{for some } \delta > 0 \textit{ and for all } i, j \in \mathbb{N}. \tag{94}$$

Then the function

$$f(\boldsymbol{x}) = \sum_{n=1}^{\infty} \frac{1}{n}\, k(\boldsymbol{x}, \boldsymbol{x}_n), \tag{95}$$

with the Gaussian kernel

$$k(\boldsymbol{x}, \boldsymbol{y}) = \exp\Big(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\sigma^2}\Big), \tag{96}$$

has finite RKHS norm

$$\|f\|_{\mathcal{H}}^2 = \sum_{i,j=1}^{\infty} \frac{1}{ij}\, k(\boldsymbol{x}_i, \boldsymbol{x}_j) < \infty, \tag{97}$$

even though

$$\|\alpha\|_{\ell_1} = \sum_{n=1}^{\infty} \frac{1}{n} = \infty. \tag{98}$$

Proof We begin by splitting the double series defining the RKHS norm into diagonal and off–diagonal parts:

$$\|f\|_{\mathcal{H}}^2 = \sum_{i=1}^{\infty} \frac{1}{i^2}\, k(\boldsymbol{x}_i, \boldsymbol{x}_i) + \sum_{i \neq j} \frac{1}{ij}\, k(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{99}$$

Since $k(\boldsymbol{x}, \boldsymbol{x}) = 1$ for all $\boldsymbol{x} \in \mathbb{R}^d$, the diagonal contribution is

$$S_{\text{diag}} = \sum_{i=1}^{\infty} \frac{1}{i^2}, \tag{100}$$

which converges (indeed, $\sum_{i=1}^{\infty} \frac{1}{i^2} = \pi^2/6$).

For the off–diagonal part, define

$$S_{\text{off}} = \sum_{i \neq j} \frac{1}{ij}\, k(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{101}$$

By symmetry and non-negativity of $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, we can write

$$S_{\text{off}} = 2 \sum_{i > j} \frac{1}{ij}\, k(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{102}$$

For $i > j$, the separation condition implies

$$\|\boldsymbol{x}_i - \boldsymbol{x}_j\| \geq (i - j)\delta, \tag{103}$$

so that

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\Big(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\Big) \leq \exp\Big(-\frac{((i-j)\delta)^2}{2\sigma^2}\Big). \tag{104}$$

Setting

$$k = i - j \quad (k \geq 1) \tag{105}$$

and writing $i = j + k$, we obtain

$$S_{\text{off}} \leq 2 \sum_{k=1}^{\infty} \exp\Big(-\frac{(k\delta)^2}{2\sigma^2}\Big) \sum_{j=1}^{\infty} \frac{1}{j(j+k)}. \tag{106}$$

We now analyze the inner sum. Using partial fractions,

$$\frac{1}{j(j+k)} = \frac{1}{k}\Big(\frac{1}{j} - \frac{1}{j+k}\Big). \tag{107}$$

Thus,

$$\sum_{j=1}^{\infty} \frac{1}{j(j+k)} = \frac{1}{k} \sum_{j=1}^{\infty} \left(\frac{1}{j} - \frac{1}{j+k}\right). \tag{108}$$

The telescoping sum yields

$$H_k := \sum_{j=1}^{\infty} \left(\frac{1}{j} - \frac{1}{j+k}\right) = \sum_{j=1}^{k} \frac{1}{j}, \tag{109}$$

where H_k is also known as the kth harmonic number. Hence,

$$\sum_{j=1}^{\infty} \frac{1}{j(j+k)} = \frac{H_k}{k}. \tag{110}$$

It follows that

$$S_{\text{off}} \leq 2 \sum_{k=1}^{\infty} \exp\Big(-\frac{(k\delta)^2}{2\sigma^2}\Big) \frac{H_k}{k}. \tag{111}$$

For large k, it holds that

$$H_k = \ln k + \gamma + o(1), \tag{112}$$

where γ is the Euler–Mascheroni constant. Moreover, the Gaussian factor

$$\exp\Big(-\frac{(k\delta)^2}{2\sigma^2}\Big) \tag{113}$$

decays exponentially in k. Therefore, the series

$$\sum_{k=1}^{\infty} \exp\Big(-\frac{(k\delta)^2}{2\sigma^2}\Big) \frac{H_k}{k} \tag{114}$$

is bounded, where we note that $\frac{H_k}{k}$ is bounded from above by 1 (for all k taken sufficiently large). Combining the diagonal and off–diagonal parts, we deduce that

$$\begin{aligned}
\|f\|_{\mathcal{H}}^2 &= \sum_{i=1}^{\infty} \frac{1}{i^2}\, k(\boldsymbol{x}_i, \boldsymbol{x}_i) + \sum_{i \neq j} \frac{1}{ij}\, k(\boldsymbol{x}_i, \boldsymbol{x}_j) && (115)\\
&= S_{\text{diag}} + S_{\text{off}} && (116)\\
&\leq \sum_{i=1}^{\infty} \frac{1}{i^2} + 2 \sum_{k=1}^{\infty} \exp\Big(-\frac{(k\delta)^2}{2\sigma^2}\Big) \frac{H_k}{k} && (117)\\
&< \infty. && (118)
\end{aligned}$$

■