Appl. Comput. Harmon. Anal. 62 (2023) 194–236

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

journal homepage: www.elsevier.com/locate/acha

Understanding neural networks with reproducing kernel Banach spaces

Francesca Bartolucci[a], Ernesto De Vito[b], Lorenzo Rosasco[c,d,e], Stefano Vigogna[f,*]

[a] *SAM - Department of Mathematics, ETH Zürich, Switzerland*
[b] *MaLGa - DIMA, Università di Genova, Italy*
[c] *MaLGa - DIBRIS, Università di Genova, Italy*
[d] *CBMM - MIT, USA*
[e] *IIT, Italy*
[f] *RoMaDS - Department of Mathematics, Università degli Studi di Roma Tor Vergata, Italy*

ARTICLE INFO

Article history:
Received 26 October 2021
Received in revised form 9 June 2022
Accepted 30 August 2022
Available online 5 September 2022
Communicated by Hrushikesh Mhaskar

Keywords:
Neural networks
Reproducing kernel Banach spaces
Representer theorem
Radon transform

ABSTRACT

Characterizing the function spaces corresponding to neural networks can provide a way to understand their properties. In this paper we discuss how the theory of reproducing kernel Banach spaces can be used to tackle this challenge. In particular, we prove a representer theorem for a wide class of reproducing kernel Banach spaces that admit a suitable integral representation and include one hidden layer neural networks of possibly infinite width. Further, we show that, for a suitable class of ReLU activation functions, the norm in the corresponding reproducing kernel Banach space can be characterized in terms of the inverse Radon transform of a bounded real measure, with norm given by the total variation norm of the measure. Our analysis simplifies and extends recent results in [45,36,37].

1. Introduction

Neural networks provide a flexible and effective class of machine learning models, by recursively composing linear and nonlinear functions. The models thus obtained correspond to nonlinearly parameterized functions, and typically require non convex optimization procedures [17]. While this does not prevent good empirical performances, it makes understanding neural network properties considerably complex. Indeed, characterizing what function classes can be well represented/approximated by neural networks is a classic problem [38,3,45,36,37,18], but it is still not fully understood. Moreover, networks with large numbers of parameters are often practically successful, seemingly contradicting the idea that models should be simple to be learned from data [59,6]. This observation raises the question of in what sense the complexity of the models is explicitly or implicitly controlled. From a functional perspective, the answer corresponds

[*] Corresponding author.
E-mail addresses: francesca.bartolucci@sam.math.ethz.ch (F. Bartolucci), ernesto.devito@unige.it (E. De Vito), lorenzo.rosasco@unige.it (L. Rosasco), vigogna@mat.uniroma2.it (S. Vigogna).

https://doi.org/10.1016/j.acha.2022.08.006
1063-5203/© 2022 Elsevier Inc. All rights reserved.

to understanding what norms can be defined and controlled on the spaces of functions defined by neural networks.

Among neural networks, there is one model where the above questions become considerably more amenable to study, namely neural networks with only one hidden layer. In this case, functions can be seen to be parameterized by measures, with networks with finitely many hidden units corresponding to measures of finite support [3]. The remarkable advantage of this framework is that the parameterization in terms of measures is linear, and functional calculus considerably simplifies. This observation is at the base of the connection between neural networks and Gaussian processes [34], as well as random features [40], which allows to bring to bear the powerful machinery of reproducing kernel Hilbert spaces [1]. However, starting at least from [5,4], it is clear that norms other than Hilbertian can be defined that might better capture the inductive biases induced by neural networks. For example, for functions parameterized by absolutely continuous measures, the L^1 norm of the corresponding densities can be considered. More generally, functional norms can be defined in terms of total variations of the corresponding measures. The study in [3] provides a clear discussion on this perspective.

The extension from a Hilbert to a Banach setting opens a number of questions. We discuss two that are relevant to our study. The first one is related to the characterization of the solution of empirical minimization problems, the so-called representer theorem. It is well known that, in a Hilbert setting, minimizers always lie in a finite dimensional subspace. Each solution is a linear combination of the reproducing kernel associated to the Hilbert space evaluated at the training set points [25,26,46]. This result has immediate computational implications and is at the base of kernel methods [47]. A natural question is then how these results extend to a Banach space of functions defined by neural networks. A number of recent results tackles this question [54,37]. A main difficulty is that the Banach spaces defined by neural networks are non-reflexive, and their definition requires some care. In this context, our first contribution is that we systematically use the machinery of reproducing kernel Banach spaces [60,29] to simplify and analyze the construction of such spaces. In the Hilbert setting, feature maps and positive definite kernels can both be equivalently used to define functions spaces with the reproducing property. For non-reflexive Banach spaces, only feature maps provide a natural approach. While a reproducing kernel can be defined, it is typically neither symmetric nor positive definite. Instead, we show that, introducing appropriate feature maps, function spaces defined by neural networks can be seen to define reproducing kernel Banach spaces of functions admitting a suitable integral representation. Through this characterization and the application of a recent technical result in [8], we can immediately derive a representer theorem. This result can be contrasted to [37], and, as discussed later, allows dealing more directly with some technical issues. We note in passing that representer theorems for neural networks have different implications than analogous results in the Hilbert setting. Unlike the Hilbert setting, they do not have immediate computational consequences, but have interesting implications from a conceptual point of view. Indeed, they imply that finite networks suffice to solve empirical minimization problems. Further, they imply an upper bound on the amount of overparameterization required.

A second line of inquiry regards the characterization of the functions and the norms corresponding to neural networks. Once again, it is instructive to look at the Hilbert setting. A main example of reproducing kernel Hilbert spaces is Sobolev spaces with smoothness sufficiently high for the embedding theorem to hold. In this case, the norm in the reproducing kernel Hilbert space can be characterized in terms of a suitable pseudo-differential operator, with the associated reproducing kernel being the corresponding Green function [57]. Again, the question is whether similar characterizations can be derived for reproducing kernel Banach spaces defined by neural networks. A recent line of works shows that results in this direction can be derived when considering the rectified linear activation function (ReLU) in the network units. A first result in this direction is derived in [45] for univariate functions, and then developed in [36] for the general multivariate case. In particular, this latter paper shows that the corresponding Banach semi-norm can be characterized using the Radon transform. These results are further developed in [37], where semi-norms are defined in terms of the Radon transform in order to prove a representer theorem for one hidden layer neural networks with (generalized) ReLU activation function. In particular, the definition of the semi-norm precedes and is in function of proving the representer theorem. Here we contribute to this line of work, refining and extending such results, as well as providing different derivations. Indeed, we show that an analogous yet finer Radon characterization holds true for the reproducing kernel Banach spaces corresponding to neural networks with (generalized) ReLU activation functions. Our construction shows that the characterization of the Banach space structure is independent of the representer theorem. Moreover, our approach provides a natural norm regularizer, thus avoiding semi-norms with resulting topological issues. Using a norm instead of a semi-norm also prevents the addition of null space elements (*i.e.* polynomials) to the neural network minimizers.

The paper is organized as follows. In Section 2 we recall the main ideas and results about learning with kernels. In Section 3 we give a short introduction to reproducing kernel Banach spaces (RKBS) and their characterization in terms of feature maps. Then, we introduce a class of integral RKBS that can model one hidden layer neural networks, and establish a representer theorem for such a class in Section 3.4. In Section 4 we focus on the special case of one hidden layer neural networks with (generalized) ReLU activation function. In particular, in Section 4.2 we characterize the corresponding norm by means of the Radon transform. Section 5 contains the proofs of the main results of Section 4.2. In Sections 3.6 and 4.3 we discuss and compare our results with [37]. Finally, in Appendix A we review the theory of the Radon transform and in Appendix B we collect some variational results that we use to prove our representer theorem.

Notation. If $x, y \in \mathbb{R}^d$, $x \cdot y$ denotes their scalar product and $|x|$ denotes the Euclidean norm. The length of a multi-index $m \in \mathbb{N}^d$ is denoted by $|m| = m_1 + \ldots + m_d$. Furthermore, if $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and $m = (m_1, \ldots, m_d) \in \mathbb{N}^d$, we use the notation $x^m = x_1^{m_1} \cdots x_d^{m_d}$ and $\partial^m = \partial_x^m = \partial_{x_1}^{m_1} \ldots \partial_{x_d}^{m_d}$. We denote by S^{d-1} the unit sphere in $\mathbb{R}^d$. The dual pairing between a locally convex topological space $\mathcal{A}$ and its topological dual space $\mathcal{A}'$ is denoted by ${}_{\mathcal{A}'}\langle \cdot\,, \cdot \rangle_{\mathcal{A}}$. For simplicity, we also write the pairings without specifying the dual pair $\mathcal{A}, \mathcal{A}'$ whenever it is clear from the context. The Fourier transform $\mathcal{F}$ is defined for $\varphi \in L^1(\mathbb{R}^d)$ by

$$\mathcal{F}\varphi(\omega) = \frac{1}{(2\pi)^{d/2}} \int\limits_{\mathbb{R}^d} \varphi(x) e^{-i\,x\cdot\omega} \mathrm{d}x, \qquad \omega \in \mathbb{R}^d,$$

and it extends to $L^2(\mathbb{R}^d)$ in the usual way.

If $\mathcal{B}$ is a Banach space, we denote by $\|\cdot\|_{\mathcal{B}}$ the corresponding norm. If $\mathcal{M}$ and $\mathcal{N}$ are two subspaces of $\mathcal{B}$, we write $\mathcal{B} = \mathcal{M} + \mathcal{N}$ to mean that

$$\mathcal{B} = \{m + n \colon m \in \mathcal{M},\ n \in \mathcal{N}\}, \qquad \mathcal{M} \cap \mathcal{N} = \{0\},$$

and we denote by $P_{\mathcal{M}}$ and $P_{\mathcal{N}}$ the corresponding projections

$$P_{\mathcal{M}}, P_{\mathcal{N}} : \mathcal{B} \to \mathcal{B}, \qquad P_{\mathcal{M}}(m+n) = m, \quad P_{\mathcal{N}}(m+n) = n,$$

so that $I = P_{\mathcal{M}} + P_{\mathcal{N}}$. If $\mathcal{M}$ and $\mathcal{N}$ are two Banach spaces, we write $\mathcal{B} = \mathcal{M} \oplus \mathcal{N}$ to mean that product space $\mathcal{M} \times \mathcal{N}$ endowed with the ℓ^1-norm

$$\|m + n\|_{\mathcal{B}} = \|m\|_{\mathcal{M}} + \|n\|_{\mathcal{N}}.$$

2. Background: learning with ERM and RKHS

In this section, we provide some background useful for the developments in later sections. In particular, we recall the main ideas behind learning via empirical risk minimization (ERM) and the need of incorporating a bias in the search of a solution space. Further, we recall the basic ideas and results related to considering reproducing kernel Hilbert spaces (RKHS) as solution spaces, in particular the representer theorem and the interpretation of the bias induced by the RKHS norm.

2.1. Background: learning with ERM

The basic problem of supervised learning is to estimate a function $f : \mathcal{X} \to \mathbb{R}$ of interest, given a (training) set of input/output pairs $D = \{(x_1, y_1), \dots, (x_n, y_N)\} \subset \mathcal{X} \times \mathbb{R}$. The problem is formalized in the setting of statistical learning theory [56,14,19], by assuming that $\mathcal{X} \times \mathbb{R}$ is a probability space with distribution P and that the training set is sampled identically and independently, that is $D \sim P^N$. Then, the function of interest is the one minimizing the expected risk

$$\min_{f \in \mathcal{T}} \mathcal{L}(f), \quad \mathcal{L}(f) = \int L(y, f(x)) dP(x, y),$$

where $L : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ is a given loss function. Here, the minimization is thought over the largest space $\mathcal{T}$ over which the expected risk is defined. We note that the expected risk can be interpreted as an idealization of the notion of test error. In practice, the minimization of the expected risk is unfeasible for at least two reasons. The first one is that the measure P is known only through the training set D. The second one is that, in practice, the search of a solution needs to be restricted to some class of functions $\mathcal{H} \subset \mathcal{T}$, called hypothesis space. The natural approach is then to consider the empirical risk minimization

$$\min_{f \in \mathcal{H}} \widehat{\mathcal{L}}(f), \quad \widehat{\mathcal{L}}(f) = \frac{1}{n} \sum_{i=1}^{N} L(y_i, f(x_i)).$$

While the choice of $\mathcal{H}$ might seem as a strong restriction, there are examples of spaces such that

$$\min_{f \in \mathcal{H}} \mathcal{L}(f) = \min_{f \in \mathcal{T}} \mathcal{L}(f),$$

sometimes called universal classes of functions [51,10]. As pointed out later, function spaces used in both kernel methods and neural networks can be shown to have this property. In this case, ERM is often modified considering

$$\min_{f \in \mathcal{H}} \widehat{\mathcal{L}}(f) + J(f),$$

where $J : \mathcal{H} \to \mathbb{R}$ is a functional, called regularizer. The idea is that the regularizer should enforce a bias in the search of a solution in $\mathcal{H}$ and help finding stable solutions. Next, we discuss a classic example of hypothesis spaces and regularizers, useful in our discussion.

2.2. RKHS, representer theorem and regularizers

We next consider the hypothesis space to be a RKHS. We begin recalling a general definition of RKHS and useful equivalent characterizations.

Definition 2.1. Let $\mathcal{X}$ be a set. A *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}$ over $\mathcal{X}$ is a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ such that:

(i) as a vector space, $\mathcal{H}$ is endowed with the pointwise operations of sum and multiplication by a scalar;
(ii) for all $x \in \mathcal{X}$, there is a constant $C_x > 0$ such that

$$|f(x)| \leq C_x \|f\|_{\mathcal{H}}, \qquad \forall f \in \mathcal{H}. \tag{1}$$

The property (1) states that, for every $x \in \mathcal{X}$, the point evaluation functional $\mathrm{ev}_x : \mathcal{H} \to \mathbb{R}$, $\mathrm{ev}_x f = f(x)$, is continuous. By the Riesz representation theorem, (1) is thus equivalent to the existence, for all $x \in \mathcal{X}$, of an element $K_x \in \mathcal{H}$ such that $f(x) = \langle f, K_x \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$. This observation leads to the following more practical characterization of RKHS [1].

Proposition 2.2. *A Hilbert space $\mathcal{H}$ of functions on $\mathcal{X}$ is a RKHS if and only if there exists a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that for all $x \in \mathcal{X}$*

(i) $K(x, \cdot) \in \mathcal{H}$,
(ii) $f(x) = \langle f, K(x, \cdot) \rangle_{\mathcal{H}}, \ \forall f \in \mathcal{H}$.

The function K is called the *reproducing kernel* and item (ii) is called the *reproducing property*. It is easy to check that every reproducing kernel is symmetric and positive definite. The reproducing kernel, often just called the kernel, is a key quantity uniquely associated to each RKHS. In the following, we will see how kernels are useful to characterize the solutions of corresponding ERM problems. More generally, it is possible to prove a converse of the above result showing that each symmetric positive definite kernel can be used to define a unique RKHS [1]. Here, we omit this characterization and recall another one which is popular in machine learning.

Proposition 2.3. *A space $\mathcal{H}$ of functions on $\mathcal{X}$ is a RKHS if and only if there exist a Hilbert space $\mathcal{F}$ and a map $\phi : \mathcal{X} \to \mathcal{F}$ such that*

(i) $\mathcal{H} = \{f_w : w \in \mathcal{F}\}$ *where* $f_w(x) = \langle \phi(x), w \rangle_{\mathcal{F}}$;
(ii) $\|f\|_{\mathcal{H}} = \inf\{\|w\|_{\mathcal{F}} : w \in \mathcal{F}, f = f_w\}$.

The map ϕ is called a *feature map* and $\mathcal{F}$ a *feature space*. By the above result, each function in a RKHS can be seen as a hyperplane in the feature space. The linear parameterization of a RKHS is explicit in the above characterization.

An extension of Definition 2.1 and its equivalent characterizations will be useful in the following, while considering neural nets. It will also be useful to recall two immediate consequences. The first is a representer theorem that characterizes the solution of the ERM regularized with the squared RKHS norm.

Theorem 2.4. *Assume $\mathcal{H}$ is a RKHS and, for every $y \in \mathbb{R}$, the function $L(y, \cdot)$ is convex. Then, the problem*

$$\min_{f \in \mathcal{H}} \widehat{\mathcal{L}}(f) + \|f\|^2_{\mathcal{H}}$$

has a unique minimizer f^* such that, for all $x \in \mathcal{X}$,

$$f^*(x) = \sum_{i=1}^{n} K(x, x_i) c_i, \quad c_i \in \mathbb{R}.$$

The above result is remarkable since it implies that the minimization over an infinite dimensional space can be replaced with a finite dimensional one. Indeed, this is the key observation behind kernel methods [47]. We end this section recalling that for several reproducing kernels the nature of the regularizers induced by the corresponding squared RKHS norm can be interpreted via an equivalent characterization.

Example 2.5 *(Differential operators and Sobolev spaces).* Let $\mathcal{X} = \mathbb{R}^d$ and $k(x, x') = e^{-\|x-x'\|}$ the Laplacian kernel. Then, for $s = d/2 + 1/2$, it can be shown that

$$\|f\|^2_{\mathcal{H}} \asymp \|f\|^2_2 + \|\Delta^{s/2} f\|^2_2,$$

where $\|f\|_2 = \int |f(x)|^2 dx$, and Δ is the Laplace Beltrami operator. Through the above characterization, functions with a small RKHS norm can be seen to be more regular. Similar reasoning can also be shown to apply to other translation invariant kernels. Interestingly, for all these examples the corresponding RKHS are universal [31].

In the following we discuss the question of whether the above results apply or can be extended to neural networks, and discuss several implications.

3. RKBS of neural networks and representer theorem

In this section, we discuss how different function spaces can be associated to neural networks. In particular, we discuss how RKBS can be used towards this end, and corresponding representer theorems derived. We first recall the basic expression for neural networks with one hidden layer and illustrate the benefit of considering the limit in which the hidden layer can have infinite width.

3.1. Infinite wide neural networks are linearly parameterized over measures

As mentioned before, a main obstacle towards studying function spaces defined by neural networks is their nonlinear parameterization. Starting from a linear function $w \cdot x$, Proposition 2.3 shows how linearly parameterized nonlinear functions can be obtained applying a non linear map to the input $w \cdot \phi(x)$. In neural networks instead, a continuous nonlinear function $\sigma : \mathbb{R} \to \mathbb{R}$ is applied also to the parameters by considering $\sigma(w \cdot x)$. Indeed, this latter expression is a simplified model of a neuron. A one hidden layer neural network is a function obtained as linear combination of several neurons

$$f(x) = \sum_{k=1}^{K} \alpha_k \sigma(w_k \cdot x - b_k), \tag{2}$$

where $w_k \in \mathbb{R}^d$ and $b_k \in \mathbb{R}$ are called the weights. The above expression can be developed considering further compositions to obtain “deeper” multilayer architectures. In this paper, we restrict our attention to one hidden layer networks. In the following, we discuss how functions spaces of neural networks can be defined very generally considering an extension of RKHS, namely RKBS. We first illustrate some basic ideas, in particular a suitable reparameterization of neural networks in terms of measures.

We use the short hand notation $\rho(x, \theta) = \sigma(w \cdot x - b)$, where $\theta = (w, b)$. The key idea is to consider the limit for large K in equation (2), that is

$$\sum_{k=1}^{K} \rho(x, \theta_k) c_k \quad \mapsto \quad \int \rho(x, \theta) d\mu(\theta). \tag{3}$$

The latter expression is the limit where the hidden layer has an infinite number of neurons. Note that, if $\mu = \sum_{k=1}^{K} \delta_{\theta_k} c_k$ then

$$\int \rho(x,\theta) d\mu(\theta) = \sum_{k=1}^{K} \rho(x,\theta_k) c_k.$$

Considering the integral in equation (3) requires some care, and in the next few sections we will discuss how function spaces can be defined with the aid of RKBS. We first discuss a simplified setting to illustrate some basic intuitions.

Example 3.1 *(Compact parameter spaces and densities).* We let $\theta \in \Theta$, where the parameter Θ is a compact subset of $\mathbb{R}^d$, and restrict to measures that are absolutely continuous with respect to the Lebesgue measure $d\theta$, so that $\mu(\theta) = p(\theta)d\theta$. Then, equation (3) becomes

$$f_\mu(x) = \int \rho(x,\theta) p(\theta) d\theta. \tag{4}$$

The above expression shows how functions are linearly parameterized by measures/ densities, and it is easy to see that they form a linear space. Different structures and in particular different norms can be considered, for example $\|f\|_{\mathcal{H}} = \|p\|_{L^2(\Theta)}$ or $\|f\|_{\mathcal{B}} = \|p\|_{L^1(\Theta)}$. It can be proved [3,43] that the first choice corresponds to considering a RKHS $\mathcal{H}$ with kernel

$$K(x,x') = \int \rho(x,\theta)\rho(x',\theta) d\theta.$$

Indeed, this result is at the base of well known connections between neural networks with RKHS, and in particular random features [40], but also with Gaussian processes [34]. The norm $\|f\|_{\mathcal{B}} = \|p\|_{L^1(\Theta)}$, instead, can be shown to define a Banach space [3], and clearly $\mathcal{H} \subset \mathcal{B}$. Hence, in general, we can expect the space $\mathcal{B}$ to have better approximation properties than $\mathcal{H}$. We remark that, while surely enlightening, this setting has at least two major limitations: first, the parameter space of commonly used neural networks is never compact; second, restricting to absolutely continuous measures excludes atomic measures, and therefore (finite width) neural networks.

The above example shows that, while a connection between RKHS and neural nets is possible, going beyond a Hilbert setting might be needed depending on the kind of structures we consider on the function space of neural networks. The fact that Banach spaces of neural networks are larger function spaces suggests that it could be interesting to explore this setting. Interestingly, recent results also suggest that the gradient descent training of neural networks might be controlling implicitly the norm in $\mathcal{B}$ [12]. Indeed, we will show next that certain Banach spaces are naturally associated to neural networks. Towards this end, we first recall the basic facts about RKBS.

3.2. Reproducing kernel Banach spaces

Since [60], several definitions of RKBS have been proposed. Here, we adopt a fairly minimal definition, and refer to [29] for a comprehensive overview. Among all possible equivalent definitions of RKHS, the one in Definition 2.1 generalizes seamlessly to the Banach case. Indeed, it suffices to replace "Hilbert" with "Banach".

Definition 3.2. Let $\mathcal{X}$ be a set. A *reproducing kernel Banach space* (RKBS) $\mathcal{B}$ over $\mathcal{X}$ is a Banach space $\mathcal{B}$ of functions $f : \mathcal{X} \to \mathbb{R}$ such that:

(i) as a vector space, $\mathcal{B}$ is endowed with the pointwise operations of sum and multiplication by a scalar;
(ii) for all $x \in \mathcal{X}$, there is a constant $C_x > 0$ such that

$$|f(x)| \leq C_x \|f\|_{\mathcal{B}}, \qquad \forall f \in \mathcal{B}. \tag{5}$$

As for RKHS, the property (5) is equivalent to the fact that for every $x \in \mathcal{X}$ there exists an element $\mathrm{ev}_x \in \mathcal{B}'$ such that

$$f(x) = {}_{\mathcal{B}'}\langle \mathrm{ev}_x, f \rangle_{\mathcal{B}}, \qquad \forall f \in \mathcal{B}. \tag{6}$$

However, unlike for RKHS, this does not lead to a natural notion of reproducing kernel, and thus to a characterization as in Proposition 2.2, because in general $\mathcal{B}$ is not isomorphic to its dual. Interestingly, the characterization of Proposition 2.3 in terms of feature maps generalizes naturally [13,29]. We report the proof for the sake of completeness.

Proposition 3.3. *A space $\mathcal{B}$ of functions on $\mathcal{X}$ is a RKBS if and only if there exist a Banach space $\mathcal{F}$ and a map $\phi : \mathcal{X} \to \mathcal{F}'$ such that*

(i) $\mathcal{B} = \{f_\mu : \mu \in \mathcal{F}\}$ *where* $f_\mu(x) = {}_{\mathcal{F}'}\langle \phi(x), \mu \rangle_{\mathcal{F}}$*;*
(ii) $\|f\|_{\mathcal{B}} = \inf\{\|\mu\|_{\mathcal{F}} : \mu \in \mathcal{F}, f = f_\mu\}$.

Proof. Let $\mathcal{B}$ be a RKBS of functions on $\mathcal{X}$. Define $\mathcal{F} = \mathcal{B}$ and the canonical feature map

$$\phi : \mathcal{X} \to \mathcal{B}', \qquad \phi(x) = \mathrm{ev}_x,$$

where ev_x is defined by (6), so that $f_\mu = \mu$ for all $\mu \in \mathcal{B}$. Then both (i) and (ii) are clear. Conversely, suppose we have a Banach space $\mathcal{F}$ and a map $\phi : \mathcal{X} \to \mathcal{F}'$, and define a vector space $\mathcal{B}$ of functions on $\mathcal{X}$ as in (i). Then, the norm in (ii) makes $\mathcal{B}$ a normed space. To see that $\mathcal{B}$ is complete, consider the linear map $\phi_* \colon \mathcal{F} \to \mathcal{B}$ given by

$$\phi_*(\mu) = f_\mu,$$

and observe that the kernel $\mathcal{N}$ of ϕ_* is a closed subspace of $\mathcal{F}$, as it is the intersection of closed subspaces in $\mathcal{F}$:

$$\mathcal{N} = \bigcap_{x \in \mathcal{X}} \operatorname{Ker} \phi(x).$$

Then, $\mathcal{F}/\mathcal{N}$ is a Banach space with respect to the norm

$$\|\pi(\mu)\|_{\mathcal{F}/\mathcal{N}} = \inf\{\|\nu\|_{\mathcal{F}} : \nu \in \mathcal{F}, \pi(\nu) = \pi(\mu)\},$$

where $\pi \colon \mathcal{F} \to \mathcal{F}/\mathcal{N}$ denotes the canonical projection [44, Chapter 1,Theorem 1.41]. By definition, $\mathcal{B}$ is isomorphic to $\mathcal{F}/\mathcal{N}$, and

$$\|f_\mu\|_{\mathcal{B}} = \inf\{\|\nu\|_{\mathcal{F}} : \nu \in \mathcal{F}, f_\nu = f_\mu\} = \|\pi(\mu)\|_{\mathcal{F}/\mathcal{N}}.$$

Therefore, $\mathcal{B}$ is a normed space isometrically isomorphic to the Banach space $\mathcal{F}/\mathcal{N}$, and consequently it is complete. Moreover, in view of (i), for every $f \in \mathcal{B}$ there exists $\mu \in \mathcal{F}$ such that $f = f_\mu$, and $|f(x)| = |f_\mu(x)| \leq \|\mu\|_{\mathcal{F}} \|\phi(x)\|_{\mathcal{F}'}$. Thus, for every $x \in \mathcal{X}$,

$$|f(x)| \leq \inf_{\mu\in\mathcal{F}, f=f_\mu} \|\mu\|_{\mathcal{F}} \|\phi(x)\|_{\mathcal{F}'} = \|f\|_{\mathcal{B}} \|\phi(x)\|_{\mathcal{F}'},$$

which shows that point evaluation is continuous on $\mathcal{B}$. □

Some comments are in order. As mentioned above, Proposition 3.3 gives a recipe to construct RKBS starting from a Banach space $\mathcal{F}$ and a map $\phi : \mathcal{X} \to \mathcal{F}'$. In analogy to RKHS, we call ϕ a *feature map* and $\mathcal{F}'$ a *feature space*. As in the Hilbert setting, we note that feature maps are in general not unique. Finally, we add a technical remark.

Remark 3.4. As proved in Proposition 3.3, the RKBS $\mathcal{B}$ is isometrically isomorphic to the quotient space $\mathcal{F}/\mathcal{N}$, where $\mathcal{N}$ is the closed subspace

$$\mathcal{N} = \{\mu \in \mathcal{F} : f_\mu(x) = 0 \quad \forall x \in \mathcal{X}\},$$

and the isometry is given by

$$W_\phi : \mathcal{F}/\mathcal{N} \to \mathcal{B}, \qquad W_\phi(\pi(\mu)) = f_\mu,$$

where $\pi(\mu)$ is the coset of μ. Since the dual of $\mathcal{F}/\mathcal{N}$ can be identified with the closed subspace

$$\mathcal{N}^\perp = \{\omega \in \mathcal{F}' : {}_{\mathcal{F}'}\langle \omega, \mu\rangle_{\mathcal{F}} = 0\, \forall \mu \in \mathcal{N}\} \subseteq \mathcal{F}',$$

then by duality $\mathcal{B}'$ is isometrically isomorphic to $\mathcal{N}^\perp$. In particular,

$$W'_\phi \, \mathrm{ev}_x = \phi(x), \qquad x \in \mathcal{X}, \tag{7}$$

where $W'_\phi : \mathcal{B}' \to \mathcal{N}^\perp$ denotes the dual map.

Next, we describe a class of RKBS parametrized by the space of bounded measures, which is a variant of an example in [3]. This RKBS is the example relevant to discuss spaces of functions defined by neural networks.

3.3. A class of integral RKBS

We fix a (Hausdorff) locally compact second countable topological space Θ, that can be seen as the parameter space. Then, we denote by $\mathcal{M}(\Theta)$ the Banach space of bounded measures defined on the Borel σ-algebra of Θ, and endow $\mathcal{M}(\Theta)$ with the total variation norm $\|\cdot\|_{\mathrm{TV}}$. Since Θ is second countable, the elements of $\mathcal{M}(\Theta)$ are finite Radon measures, and the Markov-Riesz representation theorem ensures that $\mathcal{M}(\Theta)$ can be identified with the dual of $\mathrm{C}_0(\Theta)$, the Banach space of continuous functions going to zero at infinity endowed with the sup norm $\|\cdot\|_\infty$. Then the TV norm is written as

$$\|\mu\|_{\mathrm{TV}} = \sup\{\langle \mu, \psi\rangle : \psi \in \mathrm{C}_0(\Theta), \|\psi\|_\infty \leq 1\}. \tag{8}$$

Keys to our construction are a function $\rho : \mathcal{X} \times \Theta \to \mathbb{R}$ and a measurable function $\beta : \Theta \to \mathbb{R}$ satisfying the following conditions:

(i) for all $x \in \mathcal{X}$

$$\sup_{\theta\in\Theta} |\rho(x,\theta)\beta(\theta)| = D_x < \infty, \tag{9}$$

for some $D_x > 0$;

(ii) for all $x \in \mathcal{X}$, the function $\rho(x, \cdot)$ is measurable.

Given the above definition we next define a RKBS of functions with a suitable integral representation and that can be seen to be parameterized in terms of measures on the parameter space. As discussed later this yields a direct connection with one hidden layer neural networks with possibly infinite width. Towards this end, we define the feature map

$$\phi : \mathcal{X} \to \mathcal{M}(\Theta)', \qquad {}_{\mathcal{M}(\Theta)}\langle \mu, \phi(x) \rangle_{\mathcal{M}(\Theta)'} = \int\limits_{\Theta} \rho(x,\theta)\beta(\theta)\mathrm{d}\mu(\theta),$$

which is well defined because of (9). Then, by Proposition 3.3 the feature map ϕ defines a RKBS $\mathcal{B}$ explicitly given by

$$\mathcal{B} = \{f_\mu : \mu \in \mathcal{M}(\Theta)\}, \tag{10a}$$

$$f_\mu(x) = \int\limits_{\Theta} \rho(x,\theta)\beta(\theta)\mathrm{d}\mu(\theta), \tag{10b}$$

$$\|f\|_{\mathcal{B}} = \inf\{\|\mu\|_{\mathrm{TV}} : f_\mu = f\}. \tag{10c}$$

We add several remarks. First, we comment on the nature of the functions ρ and β.

Remark 3.5 *(Reproducing kernel and activation functions).* The function ρ is a *reproducing kernel* in the sense of [29, Definition 2.1] (see [29, Section 3.4]). Clearly, it is always possible to include β in the definition of the kernel ρ. However, we prefer to regard $\{\rho(\cdot,\theta)\}_\theta$ as a family of basis functions (*e.g.* as identified by the choice of an activation function in neural networks), and β as a smoothing function needed to ensure that the integral in (10b) converges for all μ.

As we comment next, the introduction of the smoothing function is crucial.

Remark 3.6 *(Smoothing function β).* Condition (9) (with the measurability assumption) is necessary and sufficient to ensure that the integral in (10b) converges for all bounded measures, and thus that all the elements of the hypothesis space have an integral representation. In [37] β is not introduced, and in fact their Lemma 21 provides an integral representation only for rapidly decreasing measures. Then, the authors assume that such a representation extends to a bounded operator. Note, however, that the extension of an integral operator is not necessarily integral. For example, the L^2 extension of the Fourier transform does not admit an integral representation. On a related note, [55] considers hypothesis spaces with integral representation by imposing a growth condition on the integral kernel (see Theorem 3 therein). In our setting, such a kernel would correspond to the product of ρ and β. Since we need to keep ρ free of growth conditions (in order to plug in relevant examples of activation functions), we charge β with a decay condition. In particular, our strategy allows to seamlessly deal with neural networks defined by ReLU activation functions.

By choosing the measure μ having finite support, *i.e.*

$$\mu = \sum_{k=1}^{K} a_k\, \delta_{\theta_k}, \qquad a_k \in \mathbb{R}, \quad \theta_k \in \Theta,$$

where δ_θ is the Dirac measure at point θ, it follows that the elements of the form

$$f_\mu = \sum_{k=1}^{K} \alpha_k \rho(\cdot,\theta_k), \qquad \alpha_k = a_k\beta(\theta_k) \in \mathbb{R}, \quad \theta_k \in \Theta, \tag{11}$$

belong to $\mathcal{B}$. Note that the smoothing function β is included in the vector coefficient $(\alpha_1, \ldots, \alpha_K)$, so that it does not affect the dependence of the function f_μ on the input variable $x \in \mathcal{X}$. Functions as in (11) are the main ingredient of many learning algorithms, as for example kernel methods and one hidden layer neural networks, see Example 3.13 and Example 3.12 below. As observed earlier, equation (10b) provides a pointwise integral representation of the elements of $\mathcal{B}$. However, by (11), for each $\theta \in \Theta$

$$f_\theta = f_{\delta_\theta} = \rho(\cdot, \theta)\beta(\theta) \in \mathcal{B}, \qquad \|f_\theta\|_{\mathcal{B}} \leq \|\delta_\theta\|_{\mathrm{TV}} = 1, \tag{12}$$

then

$$f_\mu = \int\limits_\Theta f_\theta \, \mathrm{d}\mu(\theta), \tag{13}$$

where the integral is in the Bochner sense provided that $\theta \mapsto f_\theta$ is measurable as a map from Θ to $\mathcal{B}$. Finally, observe that (7) can be written as

$$W'_\phi \, \mathrm{ev}_x = \rho(x, \cdot)\beta \in \mathcal{M}(\Theta)'.$$

3.4. Representer theorem

We now derive a general representer theorem for the class of RKBS given by (10). As discussed next, this amounts to providing explicit characterization of the solutions to empirical minimization problems in machine learning and beyond. Following the setting described in Section 2, we consider the problem

$$\inf_{f \in \mathcal{B}} \left(\frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i)) + \|f\|_{\mathcal{B}} \right). \tag{14}$$

We are interested in the case where the hypothesis space is the RKBS given by (10) and $\|\cdot\|_{\mathcal{B}}$ is the corresponding norm. With this choice, even existence of a solution is non trivial since in general $\mathcal{B}$ is non-reflexive, so that the closed balls are not even weakly compact. In the following we establish conditions under which minimizers exist, and derive a general representer theorem.

First, we need a result showing that (14) can be reformulated as a minimization over the space of measures $\mathcal{M}(\Theta)$. The key observation is that $\mathcal{M}(\Theta)$ can be endowed with the weak* topology, with respect to which the closed balls are indeed compact.

Proposition 3.7. *Take $\rho : \mathcal{X} \times \Theta \to \mathbb{R}$, $\beta : \Theta \to \mathbb{R}$ satisfying (9), and set $\mathcal{B}$ as the corresponding RKBS defined in (10). Then*

$$\inf_{f \in \mathcal{B}} \left(\frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i)) + \|f\|_{\mathcal{B}} \right) = \inf_{\mu \in \mathcal{M}(\Theta)} \left(\frac{1}{N} \sum_{i=1}^N L(y_i, f_\mu(x_i)) + \|\mu\|_{\mathrm{TV}} \right).$$

Furthermore, if μ^* is any minimizer of

$$\inf_{\mu \in \mathcal{M}(\Theta)} \left(\frac{1}{N} \sum_{i=1}^N L(y_i, f_\mu(x_i)) + \|\mu\|_{\mathrm{TV}} \right), \tag{15}$$

then $f^* = f_{\mu^*}$ is a minimizer of problem (14).

Proof. By definition of $\mathcal{B}$, we have

$$\begin{aligned}\inf_{f\in\mathcal{B}}\left(\frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i))+\|f\|_{\mathcal{B}}\right)&=\inf_{\mu\in\mathcal{M}(\Theta)}\left(\frac{1}{N}\sum_{i=1}^{N}L(y_i,f_\mu(x_i))+\|f_\mu\|_{\mathcal{B}}\right)\\&=\inf_{\mu\in\mathcal{M}(\Theta)}\left(\frac{1}{N}\sum_{i=1}^{N}L(y_i,f_\mu(x_i))+\inf_{\substack{\nu\in\mathcal{M}\\ f_\nu=f_\mu}}\|\nu\|_{\mathrm{TV}}\right)\\&=\inf_{\substack{\mu,\nu\in\mathcal{M}(\Theta)\\ f_\nu=f_\mu}}\left(\frac{1}{N}\sum_{i=1}^{N}L(y_i,f_\mu(x_i))+\|\nu\|_{\mathrm{TV}}\right)\\&=\inf_{\nu\in\mathcal{M}(\Theta)}\left(\frac{1}{N}\sum_{i=1}^{N}L(y_i,f_\nu(x_i))+\|\nu\|_{\mathrm{TV}}\right).\end{aligned}$$

Now let assume that μ^* is a minimizer of (15). Then, for all $\nu\in\mathcal{M}(\Theta)$,

$$\left(\frac{1}{N}\sum_{i=1}^{N}L(y_i,f_{\mu^*}(x_i))+\|\mu^*\|_{\mathrm{TV}}\right)\leq\left(\frac{1}{N}\sum_{i=1}^{N}L(y_i,f_\nu(x_i))+\|\nu\|_{\mathrm{TV}}\right).$$

Fix $\mu\in\mathcal{M}(\Theta)$ and take the infimum over all ν such that $f_\nu=f_\mu$, then

$$\left(\frac{1}{N}\sum_{i=1}^{N}L(y_i,f_{\mu^*}(x_i))+\|\mu^*\|_{\mathrm{TV}}\right)\leq\left(\frac{1}{N}\sum_{i=1}^{N}L(y_i,f_\mu(x_i))+\|f_\mu\|_{\mathcal{B}}\right).$$

With the choice $\mu=\mu_*$, we have $\|\mu^*\|_{\mathrm{TV}}\leq\|f_{\mu^*}\|_{\mathcal{B}}$ and, clearly, $\|f_{\mu^*}\|_{\mathcal{B}}\leq\|\mu^*\|_{\mathrm{TV}}$, so that

$$\left(\frac{1}{N}\sum_{i=1}^{N}L(y_i,f_{\mu^*}(x_i))+\|f_{\mu^*}\|_{\mathcal{B}}\right)\leq\left(\frac{1}{N}\sum_{i=1}^{N}L(y_i,f_\mu(x_i))+\|f_\mu\|_{\mathcal{B}}\right),$$

which concludes the proof. □

The next corollary shows that the minimization problem (15) can be regarded as two nested minimization problems where the external one is over a finite-dimensional vector space. As discussed in the following, this result can be directly compared to the classic results for RKHS, highlighting similarities but also crucial differences.

Corollary 3.8. *With the setting of Proposition 3.7, let*

$$\mathcal{V}=\{\mu\in\mathcal{M}(\Theta):f_\mu(x_i)=0\ \forall i=1,\dots,N\}=\{\rho(x_1,\cdot)\beta,\dots,\rho(x_N,\cdot)\beta\}^\perp,\tag{16}$$

where the orthogonal $^\perp$ *is taken with respect to the pairing* ${}_{\mathcal{M}(\Theta)'}\langle\cdot,\cdot\rangle_{\mathcal{M}(\Theta)}$. *Then* $\mathcal{V}$ *is a closed subspace of* $\mathcal{M}(\Theta)$, *and there exists a finite-dimensional subspace* $\mathcal{W}\subset\mathcal{M}(\Theta)$ *with* $\dim\mathcal{W}\leq N$ *such that*

$$\mathcal{M}(\Theta)=\mathcal{W}+\mathcal{V},$$

and

$$\inf_{\mu\in\mathcal{M}(\Theta)}\left(\frac{1}{N}\sum_{i=1}^{N}L(y_i,f_\mu(x_i))+\|\mu\|_{\mathrm{TV}}\right)=\inf_{\nu\in\mathcal{W}}\left(\frac{1}{N}\sum_{i=1}^{N}L(y_i,f_\nu(x_i))+\inf_{\tau\in\mathcal{V}}\|\nu+\tau\|_{\mathrm{TV}}\right).\tag{17}$$

Proof. Define the map $F : \mathcal{M}(\Theta) \to \mathbb{R}$,

$$F(\mu) = \left(\frac{1}{N} \sum_{i=1}^{N} L(y_i, f_\mu(x_i)) + \|\mu\|_{\mathrm{TV}} \right).$$

By the reproducing property (10b), the linear maps

$$\mu \mapsto f_\mu(x_i), \qquad i = 1, \ldots, N,$$

are continuous. Hence, $\mathcal{V}$ is a closed subspace of $\mathcal{M}(\Theta)$ with finite co-dimension no larger than N, and therefore there is a finite dimensional subspace $\mathcal{W}$, $\dim \mathcal{W} \leq N$, such that

$$\mathcal{M}(\Theta) = \mathcal{W} + \mathcal{V}.$$

Moreover, for all $\mu = \nu + \tau$ with $\nu \in \mathcal{W}$ and $\tau \in \mathcal{V}$, we have

$$F(\mu) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f_\nu(x_i)) + \|\nu + \tau\|_{\mathrm{TV}},$$

whence (17) becomes clear. □

Corollary 3.8 is closely related to Theorem 2.4. However, there are some important differences. The existence of the finite-dimensional subspace $\mathcal{W}$ strongly depends on the fact that $\mathcal{V}$ has finite co-dimension. Moreover, in general there is not a canonical choice for the complement $\mathcal{W}$ and the total variation norm does not preserve the decomposition, *i.e.* in general $\mathcal{M}(\Theta)$ is isomorphic to $\mathcal{W} \oplus \mathcal{V}$, but the isomorphism is not an isometry. For a RKHS $\mathcal{H}$, there is a canonical choice $\mathcal{W} = \mathcal{V}^\perp$ and, for such a choice, $\|\nu + \tau\|_{\mathcal{H}}^2 = \|\nu\|_{\mathcal{H}}^2 + \|\tau\|_{\mathcal{H}}^2$, so that the inner minimization problem in (17) has $\tau = 0$ as solution. Further, since $\mathcal{M}(\Theta)$ is not reflexive, in general $\mathcal{V}$ is only weakly closed (being convex), and it is not easy to show the existence of a minimizer for the inner minimization problem.

To overcome this issue, we next strengthen condition (9) by assuming that

$$\rho(x, \cdot)\beta \in \mathrm{C}_0(\Theta), \qquad \forall x \in \mathcal{X}, \tag{18}$$

which clearly implies (9). This assumption is equivalent to assuming that the feature map

$$\phi : \mathcal{X} \to C_0(\Theta) \subset \mathcal{M}(\Theta)'$$

takes values in the pre-dual of $\mathcal{M}(\Theta)$ (compare with the assumption in [55, Theorem 1, item 2]). Moreover, for all $x \in \mathcal{X}$,

$$W_\phi' \,\mathrm{ev}_x = \rho(x, \cdot)\beta \in \mathrm{C}_0(\Theta).$$

We stress that, in many examples, given a function ρ, it is easy to find a smoothing function β such that (18) holds true without modifying the form of the solutions (11) (as functions of x). On the other hand, the choice of β does affect the norm of the solutions, albeit in a simple way. Indeed, as seen later, it simply corresponds to renormalizing the coefficients. Under condition (18), we provide a representer theorem for the RKBS defined by (10). More precisely, we show that ERM minimizers always exist, and are of the form (11). Our proof takes care of some delicate topological issues (see Remark B.4). It is based on [8, Theorem 3.3], the statement of which is given in Appendix B for the sake of completeness.

Theorem 3.9. *Assume that (18) holds true and, for every $y \in \mathbb{R}$, the function $L(y, \cdot)$ is convex and coercive in the second entry. Then, the problem*

$$\inf_{f \in \mathcal{B}} \left(\frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \|f\|_{\mathcal{B}} \right)$$

admits solutions f^* such that, for all $x \in \mathcal{X}$,

$$f^*(x) = \sum_{k=1}^{K} \alpha_k \rho(x, \theta_k), \qquad \alpha_k \in \mathbb{R} \setminus \{0\}, \quad \theta_k \in \Theta, \tag{19}$$

$$\|f^*\|_{\mathcal{B}} \leq \sum_{k=1}^{K} |\alpha_k \beta(\theta_k)^{-1}|, \tag{20}$$

with $K \leq N$ and $\beta(\theta_k) \neq 0$ for all $k = 1, \ldots, K$.

Proof. In view of Proposition 3.7 and (11), to establish (19) it is enough to consider the minimization problem (15) on the space $\mathcal{M}(\Theta)$, and show that there exists a measure μ with finite support of cardinality at most N that minimizes (15). Towards this end, we apply Theorem B.3.

We set $U = \mathcal{M}(\Theta)$ endowed with the weak* topology, so that U is a locally convex topological vector space. We define

$$\mathcal{A} : U \to \mathbb{R}^N, \qquad (\mathcal{A}\mu)_i = f_\mu(x_i) = {}_{\mathcal{M}(\Theta)'}\langle \phi(x_i), \mu \rangle_{\mathcal{M}(\Theta)} = {}_{C_0(\Theta)'}\langle \mu, \phi(x_i) \rangle_{C_0(\Theta)}.$$

By (18), $\mathcal{A}$ is a continuous linear operator from U to $\mathbb{R}^N$, regarded as Hilbert space with respect to the Euclidean scalar product. Furthermore, by assumption on L, the function

$$F : \mathbb{R}^N \to (-\infty, +\infty], \qquad F(w) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, w_i), \qquad w = (w_1, \ldots, w_N) \in \mathbb{R}^N,$$

is convex and coercive on $\mathbb{R}^N$ with domain $\mathbb{R}^N$, thus it is continuous and, hence, lower semi-continuous. We set $H = \operatorname{range} \mathcal{A}$, which is a Hilbert space since it a closed subspace of $\mathbb{R}^N$. With a slight abuse of notation, we regard F as a map defined on H and $\mathcal{A}$ as a map onto H, so that $\mathcal{A}$ becomes surjective. By (8), the total variation norm, regarded as a seminorm from U into $(-\infty, +\infty]$, is the superior envelope of lower semi-continuous functions, hence it is weakly continuous [9, page 11, item 4], its domain is U and its kernel is trivial. Furthermore, the Banach-Alaoglu theorem gives that the balls $\{\nu \in \mathcal{M}(\Theta) : \|\nu\|_{\mathrm{TV}} \leq R\}$ are weakly* compact for every $R > 0$, so that, according to the definition in [8, Assumption H1], the norm $\| \cdot \|_{\mathrm{TV}}$ is coercive on U.

By Theorem B.3, the problem (15) has minimizers of the form

$$\mu = \sum_{k=1}^{K} a_k u_k, \qquad K \leq N, \quad a_k > 0, \quad \sum_k a_k = \|\mu\|_{\mathrm{TV}}, \quad u_k \in \mathrm{Ext}(B),$$

where B is the unit ball in $\mathcal{M}(\Theta)$ and $\mathrm{Ext}(B)$ is the set of extremal points of B (see Definition B.1). Furthermore, thanks to Lemma B.2,

$$\mathrm{Ext}(B) = \{\pm \delta_\theta : \theta \in \Theta\},$$

so that μ is a measure with finite support of cardinality at most N. We thus set $f^* = f_\mu$ and

$$\alpha_k = \begin{cases} a_k\beta(\theta_k) & u_k = \delta_{\theta_k} \\ -a_k\beta(\theta_k) & u_k = -\delta_{\theta_k} \end{cases} .$$

By (11) we have $\alpha_k = a_k\beta(\theta_k) \neq 0$ if and only if $\beta(\theta_k) \neq 0$, so that (20) holds true by removing the parameters θ_k such that $\beta(\theta_k) = 0$, as a consequence of (10c) and the fact that $\sum_k a_k = \|\nu\|_{\mathrm{TV}}$. □

Remark 3.10. While our main motivation is supervised learning, and thus we focus on minimizing objectives defined by loss functions, it is clear from the working assumptions of Theorem B.3 that Theorem 3.9 holds true for more general variational problems, arising from different choices of sampling $\mathcal{A}$ and finite-data constraint F (see [8]).

Remark 3.11. The above result is close to [55, Theorem 1], [37, Theorem 1], where in both cases there is an extra polynomial term. It is also close to [8, Theorem 4.2], [54, Section 4.1], that are stated for $\mathcal{M}(\Theta)$. For further details and comparisons, see Sections 3.6 and 4.3.

3.5. Neural network RKBS

We start discussing some examples illustrating how the above results specialize to neural networks (we further develop this discussion in later sections).

Example 3.12 *(One hidden layer neural networks).* Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous (nonlinear) activation function. A one hidden layer neural network is a function

$$f(x) = \sum_{k=1}^{K} \alpha_k \sigma(w_k \cdot x - b_k), \tag{21}$$

with $w_k \in \mathbb{R}^d$ and $b_k \in \mathbb{R}$. Let $\Theta = \mathbb{R}^{d+1}$, $\rho(x, \theta) = \sigma(w \cdot x - b)$ for $\theta = (w, b)$, and pick a β satisfying (18). Applying Theorem 3.9, we obtain solutions of the form (21), with $K \leq N$. Typical examples of σ are sigmoidal functions, *i.e.* functions satisfying $\lim_{t\to-\infty} \sigma(t) = 0$ and $\lim_{t\to+\infty} \sigma(t) = 1$, and the widely used Rectified Linear Unit (ReLU) $\sigma(t) = \max\{0, t\}$. It is well known that for all these choices of σ the corresponding hypothesis classes are universal [15,38]. In Section 4 we will be studying in full detail the RKBS and corresponding norm associated with one hidden layer neural networks with (generalized) ReLU activation function.

Example 3.13 *(RBF networks & kernel mean embedding).* Assume that $\mathcal{X}$ is a compact topological space and $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a continuous semi-positive definite kernel. For $\mathcal{X} = \mathbb{R}^d$, a classic example is the Gaussian kernel $\kappa(x, x') = e^{-\|x-x'\|^2\gamma}$, which is also an example of Radial Basis Function (RBF) [39]. Let $\mathcal{H}$ be the corresponding reproducing kernel Hilbert space and $\mathcal{B}$ be the Banach space given by (10a) with the choice $\Theta = \mathcal{X}$, $\rho = \kappa$ and $\beta = 1$. Equation (12) gives that $f_x = \kappa(\cdot, x) = \kappa_x$ for all $x \in \mathcal{X}$, so that (13) becomes

$$f_\mu = \int_{\mathcal{X}} \kappa_x \, \mathrm{d}\mu(x) \in \mathcal{H}.$$

It is interesting to note that this is exactly the kernel mean embedding of μ (see [32] and references therein). Hence $\mathcal{B}$ is a subspace of $\mathcal{H}$ and, since the kernel mean embedding is continuous from $\mathcal{M}(\Theta)$ into $\mathcal{H}$, the norm $\|\cdot\|_{\mathcal{B}}$ is stronger than the norm induced by the scalar product of $\mathcal{H}$. For example, if the kernel κ is characteristic [32], the map $\mu \mapsto f_\mu$ is injective, so that $\mathcal{B}$ is isometrically isomorphic to $\mathcal{M}(\Theta)$, which is not separable, whereas $\mathcal{H}$ is separable since $\mathcal{X}$ is. Still, Theorem 3.9 states the existence of solutions of the form

$$f = \sum_{i=1}^{K} \alpha_i K_{x'_i}, \qquad x'_i \in \mathcal{X},$$

with $K \leq N$. Note however that Theorem 3.9 does not imply that the points x'_i belong to the training set $\{x_i\}_{i=1}^N$. For a related setting, see also [2].

In later sections, we will further develop the study of RKBS corresponding to neural networks defined by generalized ReLU functions and characterize their norm. Before that, we discuss the representer theorem we proved, reviewing classical as well recent related results.

3.6. Discussion: representer theorems in learning, Banach and variational theory

The representer theorem originates from the work of [25,26] on interpolation and smoothing problems in reproducing kernel Hilbert spaces, and plays a key role in kernel methods [46,47]. In a simple form, the classical representer theorem asserts that the solution of the regularized empirical risk minimization on a RKHS is a finite linear combination of the kernel evaluated at the input data points. This result is both conceptually and practically remarkable, since it allows to compute the solution of an infinite-dimensional models solving a finite dimensional problem.

In a broader sense, the representer theorem can also be seen as a sparsity result, showing the existence of solutions that are combinations of at most as many elements as the number of samples, regardless of how high the dimension of the hypothesis class is. Sparsity is an important property in machine learning (as well as in signal processing), and can be enforced by constraining the ℓ^1 norm of the model parameters [52,11]. In a finite-dimensional model, sparsity is essentially a consequence of Carathéodory's convex hull theorem (see *e.g.* [42, Section B.1]). Sparse models naturally generalize to infinite dimensions by replacing the linear coefficients with the integration with respect to a measure, and the ℓ^1 norm with the TV norm. Along these lines, [3,41] consider superpositions of infinitely many (and more than countable) features with TV regularization. [41, Theorem 1] can be seen as a representer theorem for bounded features and positive measures, based on an extension of Carathéodory's theorem to positive measures [41, Theorem 2]. Note that these constructions go beyond kernel methods and RKHS, and in particular in the direction of neural networks as described in previous sections, hence requiring different tools from functional analysis.

The approach relevant to our study is given by reproducing kernel Banach spaces. The paper [60] introduces reflexive RKBS and proves a representer theorem (Theorem 19) for minimal norm interpolation on uniformly convex RKBS (assuming linearly independent features at the sample points). A different approach is given in [13]. Uniform convexity is assumed so that the Riesz representation theorem holds, thus ensuring that continuous linear functionals are semi-inner products. Using bilinear forms instead of inner products, [50,49] handle non-reflexive spaces, and study in particular RKBS with ℓ^1 or TV norm. Their construction starts directly from a kernel function, on which they impose admissibility conditions to obtain representer theorems, see [50, Theorem 4.8, Corollary 4.9], [49, Theorem 2.4]. Non-reflexive p-norm RKBS are constructed in [58] via generalized Mercer kernels, although the representer theorems require reflexivity. Further definitions of RKBS are reviewed and unified in [29]. While the authors provide a general framework to construct RKBS and kernels by pairs of feature maps, their representer [29, Theorem 4.4] still assumes reflexivity of the feature space. We remark that even in the non-reflexive spaces considered in [50,49] the kernel is a function on the square of the input space, and therefore the model can not accommodate typical basis functions parameterized by a different parameter space than the input space, thus ruling out integral feature models [3,41] and neural networks.

The full generality of representer theorems beyond reflexive spaces can be found in optimization and variational theory, where they have come to mean virtually any result establishing the existence of sparse solutions to empirical minimization problems with convex regularization. This kind of problems has a long history. A notable example is Radon measure recovery with TV regularization, for which ante litteram representer theorems (for bounded domains) can be found in [16,61], stating the existence of solutions that are finite linear combinations of Dirac deltas. The proof of these results is crucially based on the Krein–Milman theorem and the characterization of extremal points. A more general setting has been recently developed in [55]. Here, the authors start from a pseudo-differential operator L, and consider the inverse problem over an associated native space $\mathcal{M}_{\mathrm{L}}$ of functions on $\mathbb{R}^d$ with generalized TV seminorm $\| \mathrm{L} \cdot \|_{\mathrm{TV}}$. Then, they show that the extremal points of such a problem are L-splines, *i.e.* functions which are sparsified by L, plus a term in the (finite-dimensional) kernel of L. This point of view has been considered by [37] and extended from $\mathbb{R}^d$ to $\mathbb{P}^d$ with the notion of ridge spline, of which ReLU neural networks are examples. The papers [7,8] introduce an extremely general variational framework that extends [55] to inverse problems on locally convex spaces with abstract convex [7] or seminorm [8] regularization. The corresponding representers are established: [7, Theorem 1] assumes a priori the existence of minimizers and focuses on the geometry of the solution set, whereas [8, Theorem 3.3] provides sufficient topological conditions for the existence of minimizers.

In summary, we can roughly identify three lines of work studying representer theorems: representers for learning models (classically kernel methods, more recently neural networks), representers for RKBS (generalizing RKHS), and representers in variational theory. Recently, the abstract variational framework has been applied and reconnected to machine learning. The paper [54] proves a general representer theorem for dual pairs of Banach spaces, which can be specialized to a wide range of learning problems, including sparse regularization on non-reflexive spaces (using [7, Theorem 1]). In [37], [8, Theorem 4.2] is applied to provide a representer theorem for neural networks with ReLU (type) activation function. In our paper, we further incorporate and exploit the ingredient of (non-reflexive) RKBS. While the RKBS structure is implicitly present in several previous works [42,3,37], its role in the explicit construction and characterization of neural network models was not clear or emphasized. In our work, we show how such a structure allows to directly derive representer theorems for feature models and neural networks from general variational theory. For a detailed comparison between our results and [37] we refer to Section 4.3.

4. Banach representation and Radon regularization of ReLU neural networks

In this section we discuss the RKBS associated with truncated power activation functions, including the ReLU. This is related to the results in [37], but here we follow a dual approach and provide a finer characterization. First, we define a hypothesis space $\mathcal{B}_m$ as a RKBS parametrized by $\mathcal{M}(\Theta)$ for a suitable choice of Θ and $\rho = \rho_m$. Then, we characterize the norm of $\mathcal{B}_m$ by means of the Radon transform.

4.1. The hypothesis space

Let S^{d-1} be the unit sphere in $\mathbb{R}^d$, and let

$$\Xi = S^{d-1} \times \mathbb{R}$$

with the product topology, which makes it a locally compact second countable space. Given $\mu \in \mathcal{M}(\Xi)$, we set $\mu^\vee \in \mathcal{M}(\Xi)$ to be the bounded measure defined by

$$\mu^\vee(E) = \mu(-E)$$

for every Borel set $E \subset \Xi$. We define the subspaces of even and odd measures as

$$\mathcal{M}(\Xi)_{\mathrm{even}} = \{\mu \in \mathcal{M}(\Xi) : \mu^\vee = \mu\},$$
$$\mathcal{M}(\Xi)_{\mathrm{odd}} = \{\mu \in \mathcal{M}(\Xi) : \mu^\vee = -\mu\}.$$

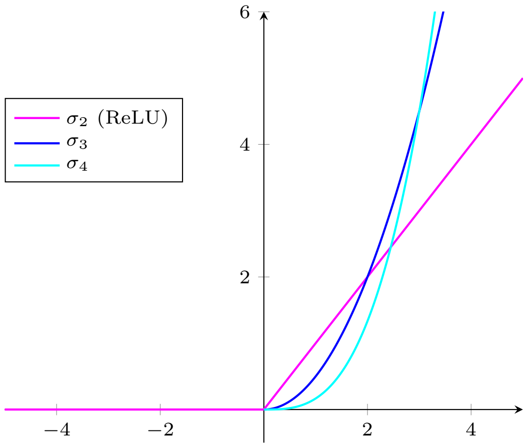

Fig. 1. ReLU-type activation functions: the ReLU σ_2, and the truncated power functions σ_3 and σ_4.

Furthermore, for every $\mu \in \mathcal{M}(\Xi)$, we define the even and odd part of μ as

$$\mu_{\text{even}} = \frac{\mu + \mu^{\vee}}{2} \in \mathcal{M}(\Xi)_{\text{even}}, \qquad \mu_{\text{odd}} = \frac{\mu - \mu^{\vee}}{2} \in \mathcal{M}(\Xi)_{\text{odd}}.$$

Every $\mu \in \mathcal{M}(\Xi)$ can be written as the sum $\mu = \mu_{\text{even}} + \mu_{\text{odd}}$ and this factorization is unique, so that

$$\mathcal{M}(\Xi) = \mathcal{M}(\Xi)_{\text{even}} + \mathcal{M}(\Xi)_{\text{odd}}.$$

Moreover, for every integer $m \geq 2$, we define the truncated power activation function $\sigma_m \colon \mathbb{R} \to \mathbb{R}$ as

$$\sigma_m(t) = \frac{1}{(m-1)!} \max\{0, t\}^{m-1}, \qquad t \in \mathbb{R} \tag{22}$$

(see Fig. 1), and the correspondingly

$$\rho_m : \mathbb{R}^d \times \Xi \to \mathbb{R}, \qquad \rho_m(x, n, t) = \sigma_m(n \cdot x - t). \tag{23}$$

Note that, for $m = 2$, σ_2 corresponds to the Rectified Linear Unit (ReLU).

We choose $\beta \in C_0(\Xi)$ such that

$$\beta(n, t) > 0, \qquad \forall (n, t) \in \Xi, \tag{24a}$$

$$\beta(-n, -t) = \beta(n, t), \qquad \forall (n, t) \in \Xi, \tag{24b}$$

$$\lim_{t \to \pm\infty} (|x| + |t|)^{m-1} \sup_{n \in S^{d-1}} \beta(n, t) = 0, \qquad \forall x \in \mathbb{R}^d. \tag{24c}$$

The positivity condition (24a) is posed to characterize the kernel of the RKBS parametrization $\mu \mapsto f_\mu$ (see Lemma 5.8). The symmetry requirement (24b) allows to control the parity when dealing with Radon transform and measures (see Lemma 5.6 and Remark 5.7). The requirement (24c) ensures that condition (18) holds true (see Remark 3.6), since

$$\sup_{n \in S^{d-1}} \rho_m(x, n, t) \leq \frac{1}{(m-1)!} (|x| + |t|)^{m-1}. \tag{25}$$

An example of β satisfying the above conditions is

$$\beta(n,t) = \frac{1}{1+|t|^m}.$$

According to the framework of Section 3.3, with the choice of $\mathcal{X} = \mathbb{R}^d$ as input space and $\Theta = \Xi$ as parameter space, we define $\mathcal{B}_m$ as the RKBS with kernel ρ_m and smoothing function β, *i.e.*

$$\mathcal{B}_m = \{f_\mu : \mu \in \mathcal{M}(\Xi)\}, \tag{26a}$$

$$f_\mu(x) = \int\limits_{\Xi} \sigma_m(n\cdot x - t)\beta(n,t)\, \mathrm{d}\mu(n,t), \tag{26b}$$

$$\|f\|_{\mathcal{B}_m} = \inf\{\|\mu\|_{\mathrm{TV}} : \mu \in \mathcal{M}(\Xi), f = f_\mu\}. \tag{26c}$$

4.2. The regularization norm

The next theorem provides an alternative characterization of the norm (26c) by means of the Radon transform. A similar result is stated in [37], within a different framework. To state our result, we first need to specify a few operators. We list them here, and we refer to Appendix A for all the details. The operator $\mathcal{R}$ denotes the Radon transform from the space $\mathcal{S}_0'(\mathbb{R}^d)$ of Lizorkin distributions on $\mathbb{R}^d$ onto the space $\mathcal{S}_0'(\Xi)$ of Lizorkin distributions on the space Ξ (Definitions A.3 and A.8). The operator Λ^{d-1} is the Fourier multiplier defined by (62) and (66), and it is at the root of the inversion formulae for the Radon transform (Theorem A.9 and Corollary A.11). The operator ∂_t is the distributional derivative acting on the variable t defined in Proposition 5.2.

Theorem 4.1. *Fix an integer $m \geq 2$. Set $\mathcal{B}_m$ as the reproducing kernel Banach space with ρ_m as in (22), (23) and β satisfying (24), and let $\mathcal{Q}_m$ and $\mathcal{P}_m$ be the subspaces defined by*

$$\mathcal{Q}_m = \{f_\tau \in \mathcal{B}_m \colon \tau \in \mathcal{M}(\Xi),\ \tau^\vee = (-1)^m \tau\},$$
$$\mathcal{P}_m = \{f_\nu \in \mathcal{B}_m \colon \nu \in \mathcal{M}(\Xi),\ \nu^\vee = (-1)^{m+1} \nu\}.$$

Then $\mathcal{Q}_m$ and $\mathcal{P}_m$ are closed subspaces of $\mathcal{B}_m$ such that

$$\mathcal{B}_m = \mathcal{Q}_m + \mathcal{P}_m,$$

and

$$\mathcal{P}_m = \{p : \mathbb{R}^d \to \mathbb{R} \colon p \text{ is a polynomial of degree at most } m-1\}.$$

Moreover:

(i) *the elements $f \in \mathcal{B}_m$ are continuous functions satisfying the growth condition*

$$|f(x)| \leq C_f (1+|x|)^{m-1}, \qquad x \in \mathbb{R}^d, \tag{27}$$

so that $f \in \mathcal{S}_0'(\mathbb{R}^d)$;

(ii) *for all $\mu \in \mathcal{M}(\Xi)$, setting*

$$\tau = \frac{\mu + (-1)^m \mu^\vee}{2}, \qquad \nu = \frac{\mu + (-1)^{m+1}\mu^\vee}{2}, \tag{28}$$

we have

$$P_{\mathcal{Q}_m} f_\mu = f_\tau, \qquad P_{\mathcal{P}_m} f_\mu = f_\nu,$$

and

$$\frac{1}{2(2\pi)^{d-1}\beta} \partial_t^m \Lambda^{d-1} \mathcal{R} f_\mu = \tau; \tag{29}$$

(iii) *for all* $f \in \mathcal{B}_m$,

$$\|f\|_{\mathcal{B}_m} \leq \|P_{\mathcal{Q}_m} f\|_{\mathcal{B}_m} + \|P_{\mathcal{P}_m} f\|_{\mathcal{B}_m} \leq 2\|f\|_{\mathcal{B}_m}, \tag{30}$$

$$\|P_{\mathcal{Q}_m} f\|_{\mathcal{B}_m} = \|\frac{1}{2(2\pi)^{d-1}\beta} \partial_t^m \Lambda^{d-1} \mathcal{R} f\|_{\mathrm{TV}}, \tag{31}$$

$$\|P_{\mathcal{P}_m} f\|_{\mathcal{B}_m} = \inf\{\|\nu\|_{\mathrm{TV}} \colon \nu \in \mathcal{M}(\Xi),\, \nu^\vee = (-1)^{m+1}\nu,\, f_\nu = P_{\mathcal{P}_m} f\}; \tag{32}$$

(iv) *take a tempered distribution* $T \in \mathcal{S}'(\mathbb{R}^d)$ *such that*

$$\tau = \frac{1}{2(2\pi)^{d-1}\beta} \partial_t^m \Lambda^{d-1} \mathcal{R} T \in \mathcal{M}(\Xi), \tag{33}$$

$$T - f_\tau \in \mathcal{P}_m \tag{34}$$

then $T \in \mathcal{B}_m$ *and*

$$P_{\mathcal{Q}_m} T = f_\tau, \qquad P_{\mathcal{P}_m} = f_\nu,$$

for some $\nu \in \mathcal{M}(\Theta)$ *such that* $\nu^\vee = (-1)^{m+1}\nu$.

The proof of Theorem 4.1 is given in Section 5. Here we add some comments. Assume that m is even, in particular $m = 2$ for the ReLU (for odd m, simply interchange "even" and "odd" in what follows). The measures τ and ν defined by (28) are the even and odd parts of μ and Theorem 4.1 states that

$$\mathcal{B}_m = \{f_\tau \colon \tau \in \mathcal{M}(\Xi)_{\mathrm{even}}\} + \{f_\nu \colon \nu \in \mathcal{M}(\Xi)_{\mathrm{odd}}\}, \tag{35}$$

so that any $f \in \mathcal{B}_m$ admits a unique decomposition $f = f_\tau + f_\nu$ with $\tau \in \mathcal{M}(\Xi)_{\mathrm{even}}$ and $\nu \in \mathcal{M}(\Xi)_{\mathrm{odd}}$. The even part τ is uniquely determined by the Radon transform of f via (29), and $\|f_\tau\|_{\mathcal{B}_m} = \|\tau\|_{\mathrm{TV}}$, so that $\mathcal{Q}_m$ is isometrically isomorphic to $\mathcal{M}(\Xi)_{\mathrm{even}}$. The odd part ν over-parametrizes the finite-dimensional space $\mathcal{P}_m$ of polynomials of degree less than m and, in particular, $\|f_\nu\|_{\mathcal{B}_m} \leq \|\nu\|_{\mathrm{TV}}$. Finally, let $L = \dim(\mathcal{P}_m)$, and let $p_1, \ldots, p_L$ be an algebraic basis of $\mathcal{P}_m$. Since L is finite-dimensional, there exists a dual family $q_1, \ldots, q_L$ in $\mathcal{B}'_m$ such that

$${}_{\mathcal{B}'_m}\langle q_\ell, p_{\ell'} \rangle_{\mathcal{B}_m} = \delta_{\ell,\ell'}.$$

Then, for all $f \in \mathcal{B}_m$,

$$f_\nu = \sum_{\ell=1}^{L} {}_{\mathcal{B}'_m}\langle q_\ell, f \rangle_{\mathcal{B}_m}\, p_\ell.$$

Item (iv) provides an equivalent characterization of $\mathcal{B}_m$ as a subspace of the space of distributions, as it happens for Besov spaces [53], and it is closely related to the original approach in [37,55]. Equation (33) means that there exists a bounded measure $\tau \in \mathcal{M}(\Xi)_{\mathrm{even}}$ such that

$$\frac{1}{2(2\pi)^{d-1}}\partial_t^m \Lambda^{d-1}\mathcal{R}T = \beta\tau \qquad \text{in } \mathcal{S}_0'(\Xi).$$

Thus, $f_\tau \in \mathcal{Q}_m \subset \mathcal{B}_m \subset \mathcal{S}'(\mathbb{R}^d)$, and (34) is equivalent to assuming that the remainder $T - f_\tau$ is a polynomial of degree less than m. Without assuming (34) we have the following result, whose proof is postponed to Section 5.

Corollary 4.2. *Take a tempered distribution $T \in \mathcal{S}'(\mathbb{R}^d)$ such that (33) holds true. Then there exist a unique $f \in \mathcal{Q}_m$ and a unique polynomial p such that $T = f + p$.*

In [37,55], the polynomial degree is enforced to be smaller than m by requiring that T is a distribution satisfying the growth condition (27). Note that $\mathcal{B}_m$ satisfies (27) by construction.

Finally, we note that Theorem 3.9 immediately gives the following representer theorem.

Corollary 4.3. *Assume that, for every $y \in \mathbb{R}$, the loss function $L(y,\cdot)$ is convex and coercive in the second entry, and set $\mathcal{B}_m$ as in Theorem 4.1. Then, the problem*

$$\inf_{f\in\mathcal{B}_m}\left(\frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) + \|f\|_{\mathcal{B}_m}\right) \tag{36}$$

always has minimizers of the form

$$f(x) = \sum_{k=1}^{K} \alpha_k \sigma_m(n_k \cdot x - t_k), \tag{37}$$

where $K \le N$, $(n_k, t_k) \in S^{d-1}\times\mathbb{R}$, $\alpha_k \in \mathbb{R}\setminus\{0\}$ and

$$\|f\|_{\mathcal{B}_m} \le \sum_{k=1}^{K} |\alpha_k| \beta(n_k,t_k)^{-1}.$$

Remark 4.4. As already observed in [37, Lemma 25], the Radon regularization corresponds to several forms of coefficient regularization, such as ℓ^1-path-norm [35] and weight decay [28]. Indeed, if we take $f \in \mathcal{B}_m$ of the form

$$f(x) = \sum_{k=1}^{K} \alpha_k \sigma_m(n_k \cdot x - t_k), \tag{38}$$

where $K \in \mathbb{N}$, $(n_k, t_k) \in S^{d-1}\times\mathbb{R}$, $\alpha_k \in \mathbb{R}\setminus\{0\}$, a simple computation gives that

$$\|P_{\mathcal{Q}_m} f\|_{\mathcal{B}_m} = \|\frac{1}{2(2\pi)^{d-1}\beta}\partial_t^m \Lambda^{d-1}\mathcal{R}f\|_{\mathrm{TV}} = \sum_{k=1}^{K} |\alpha_k| \beta(n_k,t_k)^{-1}.$$

The proof follows directly by Lemma 5.6 together with the fact that

$$f = f_\mu, \qquad \mu = \sum_{k=1}^{K} \alpha_k \beta(n_k,t_k)^{-1}\delta_{(n_k,t_k)},$$

and

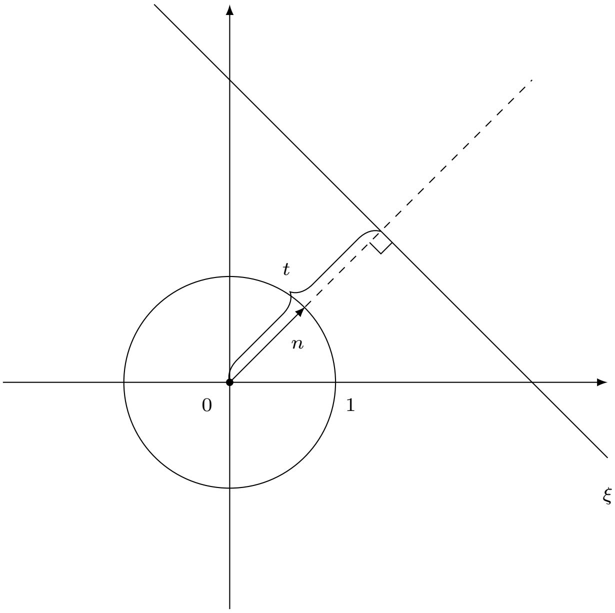

Fig. 2. The hyperplane ξ with equation $n \cdot x = t$ (two-dimensional case).

$$\left\| \frac{\mu + (-1)^m \mu^\vee}{2} \right\|_{\mathrm{TV}} = \sum_{k=1}^{K} |\alpha_k| \beta(n_k, t_k)^{-1}.$$

In the next section we provide an alternative construction of RKBS for ReLU type neural networks where the polynomial space $\mathcal{P}_m$ is avoided.

4.2.1. An alternative construction

As $\Theta = \mathbb{P}^d$, the space of all hyperplanes in $\mathbb{R}^d$, which is the natural domain of the Radon transform. For every hyperplane $\xi \in \mathbb{P}^d$ there exists $(n, t) \in \Xi$ such that

$$x \in \xi \iff x \cdot n = t.$$

see Fig. 2. The space Ξ is a double cover of $\mathbb{P}^d$ with covering map[1]

$$\Psi \colon \Xi \to \mathbb{P}^d, \qquad \Psi(n, t) = \{x \in \mathbb{R}^d : x \cdot n = t\},$$

and $\Psi(n, t) = \Psi(n', t')$ if and only if $(n', t') = (-n, -t)$. Therefore, we can identify $\mathbb{P}^d$ with the quotient space $\Xi/\sim$, where $\sim$ is the equivalence relation on Ξ given by

$$(n, t) \sim (n', t') \iff (n', t') = (-n, -t). \tag{39}$$

We denote by $[(n, t)] \in \mathbb{P}^d$ the equivalence class of $(n, t) \in \Xi$. Note that ρ_m given in (23) is not well-defined on $\mathbb{P}^d$ since $\rho_m(x, n, t) \neq \rho_m(x, -n, -t)$. To overcome this problem, we fix a measurable section

$$s \colon \mathbb{P}^d \to \Xi, \qquad s(\xi) = (n(\xi), t(\xi)),$$

[1] A double cover of a topological space X is a topological space C together with a continuous surjective map $p : C \to X$, called covering map, such that, for every $x \in X$, there exists an open neighborhood U of x such that $p^{-1}(U)$ is the union of two disjoint open sets in C, each of which homeomorphic to U via p.

i.e. s is a measurable map satisfying

$$\xi = [s(\xi)],$$

for every $\xi \in \mathbb{P}^d$. Then, we define the feature map

$$\widetilde{\phi}_m : \mathbb{R}^d \to C_0(\mathbb{P}^d) \subset \mathcal{M}(\mathbb{P}^d)'$$

given, for every $x \in \mathbb{R}^d$ and $\xi \in \mathbb{P}^d$, by

$$\widetilde{\phi}_m(x)(\xi) = \sigma_m(n(\xi) \cdot x - t(\xi))\beta(n(\xi), t(\xi)),$$

where the smoothing function β satisfies (18) and is strictly positive. Further, we suppose β to be an even function if m is even and an odd function if m is odd. This last assumption ensures that the right-hand side in formula (40) has the right parity (cf. Remark 5.7). We thus define the RKBS $\widetilde{\mathcal{B}}_m$ as the RKBS associated with the feature map $\widetilde{\phi}_m$ according to Proposition 3.3. As we will see, a crucial point to characterize the norm of $\mathcal{B}_m$ lies in Lemma 5.6. For the corresponding characterization in the space $\widetilde{\mathcal{B}}_m$, one can prove an alternative version of Lemma 5.6.

Lemma 4.5. *For every $f_\mu \in \widetilde{\mathcal{B}}_m$,*

$$\frac{1}{2(2\pi)^{d-1}} \partial_t^m \Lambda^{d-1} \mathcal{R} f_\mu = \beta\mu, \tag{40}$$

where the equality holds in $\mathcal{S}_0'(\Xi)$.

We skip the proof of Lemma 4.5 since it is similar to the proof of Lemma 5.6. Then, one can prove the following result.

Corollary 4.6. *The map $\mu \mapsto f_\mu$ is an isometry from $\mathcal{M}(\mathbb{P}^d)$ onto $\widetilde{\mathcal{B}}_m$, and*

$$\|f_\mu\|_{\widetilde{\mathcal{B}}_m} = \|\mu\|_{\mathrm{TV}} = \left\| \frac{1}{2(2\pi)^{d-1}\beta} \partial_t^m \Lambda^{d-1} \mathcal{R} f_\mu \right\|_{\mathrm{TV}}, \qquad \mu \in \mathcal{M}(\mathbb{P}^d).$$

Corollary 4.6 follows from (26c) and the fact that the map $\mu \mapsto f_\mu$ is injective. The injectivity is a consequence of Lemma 4.5, together with the fact that $\beta\mu = 0$ in $\mathcal{S}_0'(\Xi)$ implies $\mu = 0$ in $\mathcal{M}(\mathbb{P}^d)$, see Lemma 5.4. In other words, taking the feature map with values in $\mathcal{M}(\mathbb{P}^d)'$ avoids the redundant parametrization of the RKBS caused by the odd measures.

Corollary 4.7. *The problem*

$$\inf_{f \in \widetilde{\mathcal{B}}_m} \left(\frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \|f\|_{\widetilde{\mathcal{B}}_m} \right)$$

always has minimizers of the form

$$f(x) = \sum_{k=1}^{K} \alpha_k \sigma_m(n_k \cdot x - t_k),$$

where $K \le N$, $(n_k, t_k) \in S^{d-1} \times \mathbb{R}$, $\alpha_k \in \mathbb{R} \setminus \{0\}$ and

$$\|f\|_{\widetilde{\mathcal{B}}_m} = \sum_{k=1}^{K} |\alpha_k| \beta(n_k, t_k)^{-1}.$$

Remark 4.8. The introduction of the section s is technically crucial. A natural alternative to make the feature map well defined on $\mathbb{P}^d$ is to symmetrize the feature map, *i.e.*

$$\widetilde{\phi}_m(x)(\xi) = \frac{\sigma_m(n \cdot x - t)\beta(n,t) + \sigma_m(-n \cdot x + t)\beta(-n,-t)}{2}, \qquad \xi = [(n,t)].$$

However, this would result in a representation with symmetrized activation functions. For instance, for $m = 2$ we would obtain neural networks with absolute value activation function instead of the ReLU, *i.e.*

$$f(x) = \sum_{k=1}^{K} \alpha_k |n_k \cdot x - t_k|,$$

since

$$\sigma_m(n \cdot x - t) + \sigma_m(-n \cdot x + t) = |n \cdot x - t|.$$

This is roughly the strategy followed in [37], where the authors obtain representations with symmetrized activation functions, but with an additional polynomial term (see [37, Definition 5 with Remarks 6 and 7]). Note that

$$\sigma_m(n \cdot x - t) - \sigma_m(-n \cdot x + t) = -n \cdot x + t,$$

which is a polynomial of degree 1 in x. In this view, the use of the section s provides a more transparent construction.

Remark 4.9. In Corollary 4.7 we obtain the same representation as in Corollary 4.3, but with a simplified regularization compared to Theorem 4.1. Moreover, the norm of a solution f_μ is equal to (and not only controlled by) the ℓ^1 norm of the representation coefficients.

4.3. Discussion: a comparison with previous results

In [37] the authors build a family of function spaces $\mathcal{F}_m$, and seminorms $\phi_m \colon \mathcal{F}_m \to \mathbb{R}_+$ in terms of the Radon transform, such that the minimization problem

$$\inf_{f \in \mathcal{F}_m} \left(\frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \phi_m(f) \right)$$

always has minimizers of the form

$$f(x) = \sum_{k=1}^{K} \alpha_k (\sigma_m(n_k \cdot x - t_k) + (-1)^m \sigma_m(-n_k \cdot x + t_k)) + p(x), \tag{41}$$

where $K \leq N$, $(n_k, t_k) \in S^{d-1} \times \mathbb{R}$, $\alpha_k \in \mathbb{R} \setminus \{0\}$ and p is a polynomial of order less than m. We refer to Theorem 1 in [37] for the precise statement. If we compare equations (37) and (41), we can highlight two main differences. The first one consists in getting rid of the polynomial term by considering a norm, instead of a seminorm, as regularization term. The second one is to avoid solutions with symmetrized activation

Table 1
Domains (S^{d-1} denotes the unit sphere in $\mathbb{R}^d$).

$\mathbb{R}^d$	input space
$\Xi = S^{d-1} \times \mathbb{R}$	parameter space

Table 2
Function and distribution spaces ($X = \mathbb{R}^d, \Xi$). Subscripts $(\Xi)_{\mathrm{even}}$ and $(\Xi)_{\mathrm{odd}}$ denote the corresponding subspaces of even and odd measures/functions/distributions, respectively.

$\mathcal{M}(X)$	real bounded measures on X
$\mathcal{S}(X)$	Schwartz space of rapidly decreasing functions on X
$\mathcal{S}'(X)$	tempered distributions on X
$\mathcal{S}_0(X)$	Lizorkin test functions on X
$\mathcal{S}_0'(X)$	Lizorkin distributions on X

Table 3
Operators.

Radon transform	$\begin{tikzcd} \mathcal{S}_0(\mathbb{R}^d) \arrow[r, shift left, "\mathcal{R}"] \arrow[d, hook] & \mathcal{S}_0(\Xi)_{\mathrm{even}} \arrow[l, shift left, "\mathcal{R}^*"] \arrow[d, hook] \\ \mathcal{S}_0'(\mathbb{R}^d) \arrow[r, shift left, "\mathcal{R}"] & \mathcal{S}_0'(\Xi)_{\mathrm{even}} \arrow[l, shift left, "\mathcal{R}^*"] \end{tikzcd}$
Ramp filter	$\begin{tikzcd} \mathcal{S}_0(\Xi)_{\mathrm{even}} \arrow[r, "\Lambda^{d-1}"] \arrow[d, hook] & \mathcal{S}_0(\Xi)_{\mathrm{even}} \arrow[d, hook] \\ \mathcal{S}_0'(\Xi)_{\mathrm{even}} \arrow[r, "\Lambda^{d-1}"] & \mathcal{S}_0'(\Xi)_{\mathrm{even}} \end{tikzcd}$

functions as in (41). In particular, we choose the feature map with values either in $\mathcal{M}(S^{d-1} \times \mathbb{R})'$, or in $\mathcal{M}(\mathbb{P}^d)'$ but pre-composing the feature map with a measurable section $s\colon \mathbb{P}^d \to S^{d-1} \times \mathbb{R}$ (see Section 4.2.1 for full details). In view of Theorem 3.9, we first define the hypothesis space as a RKBS. Then, we show an alternative approach to rigorously characterize the regularization term, and consequently the hypothesis space, in terms of the Radon transform, which is the content of Theorem 4.1. Conversely, in [37] the authors start building *ad hoc* a family of seminorms in terms of the Radon transform, and consequently a family of hypothesis spaces. Their construction is motivated by Lemma 5.6. Then, in a second moment, they show the Banach space structure of the hypothesis spaces. A limitation of the approach in [37] is that from their construction it is not evident how to identify new hypothesis spaces for other types of activation functions. In our approach, the identification of the hypothesis space follows straightforwardly by Theorem 3.9, and it is independent of the relation between the Radon transform and the truncated power activation functions. Finally, it is worth observing that our approach provides an integral representation for all the elements of the hypothesis space. This latter result is achieved by introducing the smoothing regularizer β, that ensures the convergence of the integral (26b) without modifying the desired form for the minimizers (37). In previous works, where β is not introduced, the authors need to require alternative assumptions, as discussed in Remark 3.6.

5. Proofs of Section 4.2

We provide a detailed analysis of the main results of Section 4.2. We will make use of the classical function and distribution spaces listed in Table 2, on the domains listed in Table 1. In Table 3 we recall the main linear operators involved. For definitions and properties we refer to Appendix A.

The first lemma allows to regard $\mathcal{B}_m$ as a subspace of the space of tempered distributions. We denote by $H : \mathbb{R} \to \mathbb{R}$ the Heaviside step function

$$H(t) = \begin{cases} 0 & t < 0 \\ 1 & t \geq 1 \end{cases},$$

regarded as a temperated distribution.

Lemma 5.1. *With the above notation,*

(i) $\sigma_m \in \mathcal{S}'(\mathbb{R})$ *and*

$$\sigma_m^{(m-1)} = H, \tag{42}$$

where the equality holds true in $\mathcal{S}'(\mathbb{R})$*;*

(ii) *for all* $(n,t) \in \Xi$*,* $\rho_m(\cdot, n, t) \in \mathcal{S}'(\mathbb{R}^d)$*;*

(iii) *the elements* $f \in \mathcal{B}_m$ *are continuous functions satisfying the polynomial growth condition*

$$|f(x)| \leq C_f(1 + |x|)^{m-1}; \tag{43}$$

(iv) $\mathcal{B}_m \subset \mathcal{S}'(\mathbb{R}^d)$.

Proof. (i) and (ii) are clear. We prove (iii). Let $f \in \mathcal{B}_m$. By (26a), there exists $\mu \in \mathcal{B}_m$ such that

$$f(x) = \int_{\Xi} \sigma_m(n \cdot x - t)\beta(n,t)\ \mathrm{d}\mu(n,t).$$

Then, for every $x \in \mathbb{R}^d$,

$$\begin{aligned} |f_\mu(x)| &\leq \frac{1}{(m-1)!} \int_{\Xi} |\beta(n,t)||n \cdot x - t|^{m-1}\ \mathrm{d}\mu(n,t) \\ &\leq \frac{1}{(m-1)!} \int_{\Xi} (|x| + |t|)^{m-1} |\beta(n,t)| \mathrm{d}\mu(n,t) \\ &= \frac{1}{(m-1)!} \sum_{k=0}^{m-1} \binom{m-1}{k} |x|^k \int_{\Xi} |t|^{m-1-k} |\beta(n,t)| \mathrm{d}\mu(n,t), \end{aligned}$$

where the integrals converge by (24c). The right hand side is a polynomial of degree less than m, hence we obtain (43). We now prove that f is continuous. Since

$$f(x_0 + h) = \int_{\Xi} \sigma_m(n \cdot h + n \cdot x_0 - t)\beta(n,t)\ \mathrm{d}\mu(n,t),$$

it is enough to show that f is continuous at $x_0 = 0$. This is a consequence of the dominated convergence theorem, observing that, for each $(n,t) \in \Xi$, $x \mapsto \sigma_m(n \cdot x - t)\beta(n,t)$ is continuous and, by (25),

$$\sup_{|x| \leq 1} |\sigma_m(n \cdot x - t)\beta(n,t)| \leq (1 + |t|)^m |\beta(n,t)|,$$

where the right-hand side is integrable by (24c). Item (iv) is a direct consequence of (iii). □

The growth condition (43) is one starting point of the construction in [37] (see their equation (8)). Note that, in our construction, the smoothing function β allows us to prove that the elements of $\mathcal{B}_m$ are continuous functions.

We need to introduce the following operator, which provides a bounded inverse of the derivative. It was implicitly introduced in [55].

Proposition 5.2. *The operator*

$$\partial\colon \mathcal{S}_0(\mathbb{R}) \to \mathcal{S}_0(\mathbb{R}), \qquad \partial\psi(t) = \psi'(y),$$

is a continuous linear operator and, by duality, it extends to a weakly continuous operator on $\mathcal{S}_0'(\mathbb{R})$. The operator

$$\mathcal{A}\colon \mathcal{S}_0(\mathbb{R}) \to \mathcal{S}_0(\mathbb{R}), \qquad \mathcal{A}\psi(t) = \int\limits_{-\infty}^{t} \psi(s)\mathrm{d}s$$

is a continuous linear operator satisfying

$$\mathcal{A}\partial\psi = \partial\mathcal{A}\psi = \psi, \qquad \psi \in \mathcal{S}_0(\mathbb{R}). \tag{44}$$

By duality, $\mathcal{A}$ extends to a weakly continuous operator on $\mathcal{S}_0'(\mathbb{R})$ satisfying

$$\mathcal{A}\partial f = \partial\mathcal{A} f = f, \qquad f \in \mathcal{S}_0'(\mathbb{R}). \tag{45}$$

Proof. The first claim is a consequence of the fact that ∂ is a continuous linear operator from $\mathcal{S}(\mathbb{R})$ to $\mathcal{S}(\mathbb{R})$ and that the space of polynomials is stable under differentiation (see (58)). The family of seminorms on $\mathcal{S}(\mathbb{R})$ is given by (57). For $\varphi \in \mathcal{S}_0(\mathbb{R})$, we have

$$\mathcal{A}\varphi(x) = \int\limits_{-\infty}^{x} \varphi(t)\mathrm{d}t = -\int\limits_{x}^{+\infty} \varphi(t)\mathrm{d}t.$$

We show that $\mathcal{A}\varphi \in \mathcal{S}_0(\mathbb{R})$. For every $m \in \mathbb{N}$ and $x > 0$, we have

$$\begin{aligned}\langle x\rangle^m |\mathcal{A}\varphi(x)| &= |\int\limits_{x}^{+\infty} (1+x^2)^{\frac{m}{2}} \varphi(t)\mathrm{d}t| \le \int\limits_{x}^{+\infty} (1+t^2)^{\frac{m}{2}} |\varphi(t)|\mathrm{d}t \\ &\le \rho_{2m+4,0}(\varphi) \int\limits_{-\infty}^{+\infty} (1+t^2)^{\frac{m}{2}} \frac{1}{(1+t^2)^{m+2}}\mathrm{d}t < +\infty.\end{aligned}$$

Analogously, for every $m \in \mathbb{N}$ and $x < 0$, we have

$$\begin{aligned}\langle x\rangle^m |\mathcal{A}\varphi(x)| &= |\int\limits_{-\infty}^{x} (1+x^2)^{\frac{m}{2}} \varphi(t)\mathrm{d}t| \le \int\limits_{-\infty}^{x} (1+t^2)^{\frac{m}{2}} |\varphi(t)|\mathrm{d}t \\ &\le \rho_{2m+4,0}(\varphi) \int\limits_{-\infty}^{+\infty} (1+t^2)^{\frac{m}{2}} \frac{1}{(1+t^2)^{m+2}}\mathrm{d}t < +\infty.\end{aligned}$$

Thus, $\mathcal{A}\varphi$ is a well defined function, and for every $m \in \mathbb{N}$

$$\sup_{x\in\mathbb{R}} \langle x\rangle^m |\mathcal{A}\varphi(x)| \leq C\ \rho_{2m+4,0}(f) < +\infty \tag{46}$$

for some positive constant C. Furthermore, by definition, $\partial\mathcal{A}\varphi(x) = f(x)$, and thus, for every $m \in \mathbb{N}$ and $\alpha \geq 1$,

$$\sup_{x\in\mathbb{R}} \langle x\rangle^m |\partial^\alpha \mathcal{A}\varphi(x)| = \sup_{x\in\mathbb{R}} \langle x\rangle^m |\partial^{(\alpha-1)} f(x)| < +\infty. \tag{47}$$

Therefore, $\mathcal{A}\varphi \in \mathcal{S}(\mathbb{R})$. Moreover, since $f \in \mathcal{S}_0(\mathbb{R})$, for every $n \in \mathbb{N}$ we have

$$\int_{-\infty}^{+\infty} x^n \mathcal{A}\varphi(x)\mathrm{d}x = -\int_{-\infty}^{+\infty} x^{n+1}\partial\mathcal{A}\varphi(x)\mathrm{d}x = -\int_{-\infty}^{+\infty} x^{n+1} f(x)\mathrm{d}x = 0.$$

Hence, $\mathcal{A}\varphi \in \mathcal{S}_0(\mathbb{R})$. By (46) and (47) we have that, for every $m, \alpha \in \mathbb{N}$ and some constant C,

$$\rho_{m,\alpha}(\mathcal{A}\varphi) = \sup_{x\in\mathbb{R}} \langle x\rangle^m |\partial^\alpha \mathcal{A}\varphi(x)| \leq C\ \rho_{2m+4,\alpha-1}(f),$$

which shows that $\mathcal{A}\colon \mathcal{S}_0(\mathbb{R}) \to \mathcal{S}_0(\mathbb{R})$ is continuous. (44) is a direct consequence of the fundamental theorem of calculus. Since $\mathcal{A}$ is continuous, by duality $\mathcal{A}$ extends to a weakly continuous operator on $\mathcal{S}_0'(\mathbb{R})$ and (45) follows directly from (44). □

Note that the fact that ∂ has a bounded inverse strongly depends on the fact that its domain is $\mathcal{S}_0(\mathbb{R})$.

The next proposition is at the root of Theorem 4.1. It was first stated in [37, Lemma 18], by using the Radon transform $\mathcal{R}$. Here we provide an alternative proof based on the dual Radon transform $\mathcal{R}^*$.

Proposition 5.3. *For every $\varphi \in \mathcal{S}_0(\mathbb{R}^d)$ and for every $(n,t) \in \Xi$,*

$${}_{\mathcal{S}_0'(\mathbb{R}^d)}\langle \rho_m(\cdot, n, t), \varphi\rangle_{\mathcal{S}_0(\mathbb{R}^d)} = (-1)^m \beta(n,t)\mathcal{A}^m(\mathcal{R}\varphi)(n,t),$$

where $\mathcal{A}$ is the operator defined by (44) acting on $\mathcal{R}\varphi$ as a function of the only second variable.

Proof. Let $\varphi \in \mathcal{S}_0(\mathbb{R}^d)$. We can consider the function $T_\varphi\colon \Xi \to \mathbb{C}$ given by

$$T_\varphi(n,t) = \int_{\mathbb{R}^d} \sigma_m(x\cdot n - t)\varphi(x)\ \mathrm{d}x.$$

Reasoning as in the proof of Item (iii) of Lemma 5.1, it is possible to show that T_φ is a continuous function. We show that $T_\varphi \in \mathcal{S}_0'(\Xi)$. For every $(n,t) \in \Xi$,

$$\begin{aligned}|T_\varphi(n,t)| &\leq \int_{\mathbb{R}^d} |\sigma_m(n\cdot x - t)||\varphi(x)|\mathrm{d}x \\ &= \frac{1}{(m-1)!}\int_{\mathbb{R}^d} |n\cdot x - t|^{m-1}|\varphi(x)|\mathrm{d}x \\ &\leq \frac{1}{(m-1)!}\int_{\mathbb{R}^d} (|x| + |t|)^{m-1}|\varphi(x)|\mathrm{d}x\end{aligned}$$

$$= \frac{1}{(m-1)!} \sum_{k=0}^{m-1} \binom{m-1}{k} |t|^k \int_{\mathbb{R}^d} |x|^{m-1-k} |\varphi(x)| \mathrm{d}x,$$

which is a polynomial of order $m-1$ in the t variable. Now, we compute the expression of the m-th derivative of T_φ with respect to the variable t. Let $\psi \in \mathcal{S}_0(\Xi)$. Then

$$\begin{aligned}\langle \partial_t^m T_\varphi, \psi \rangle &= (-1)^m \langle T_\varphi, \partial_t^m \psi \rangle \\ &= (-1)^m \int_{\Xi} \left(\int_{\mathbb{R}^d} \sigma_m(n \cdot x - t) \varphi(x) \mathrm{d}x \right) \partial_t^m \psi(n,t) \mathrm{d}n \mathrm{d}t \\ &= (-1)^m \int_{\mathbb{R}^d} \left(\int_{S^{d-1}} \int_{\mathbb{R}} \sigma_m(n \cdot x - t) \partial_t^m \psi(n,t) \mathrm{d}t \mathrm{d}n \right) \varphi(x) \mathrm{d}x.\end{aligned}$$

Hence, by (42),

$$\begin{aligned}\langle \partial_t^m T_\varphi, \psi \rangle &= (-1)^m \int_{\mathbb{R}^d} \left(\int_{S^{d-1}} \int_{\mathbb{R}} H(n \cdot x - t) \partial_t \psi(n,t) \mathrm{d}t \mathrm{d}n \right) \varphi(x) \mathrm{d}x \\ &= (-1)^m \int_{\mathbb{R}^d} \left(\int_{S^{d-1}} \int_{-\infty}^{n \cdot x} \partial_t \psi(n,t) \mathrm{d}t \mathrm{d}n \right) \varphi(x) \mathrm{d}x \\ &= (-1)^m \int_{\mathbb{R}^d} \left(\int_{S^{d-1}} \psi(n, n \cdot x) \mathrm{d}n \right) \varphi(x) \mathrm{d}x.\end{aligned}$$

If ψ is an odd function, then $\int_{S^{d-1}} \psi(n, n \cdot x) \mathrm{d}n = 0$, so that $\langle \partial_t^m T_\varphi, \psi \rangle = 0$. Hence $\partial_t^m T_\varphi$ is an even distribution, *i.e.* $\partial_t^m T_\varphi \in \mathcal{S}_0'(\Xi)_{\text{even}}$. If ψ is an even function, *i.e.* $\psi \in \mathcal{S}_0(\Xi)_{\text{even}}$, Definition A.5 gives

$$\langle \partial_t^m T_\varphi, \psi \rangle = (-1)^m \int_{\mathbb{R}^d} \mathcal{R}^* \psi(x) \ \varphi(x) \mathrm{d}x.$$

Therefore, (61) gives that, for all $\psi \in \mathcal{S}_0(\Xi)_{\text{even}}$,

$$\langle \partial_t^m T_\varphi, \psi \rangle = (-1)^m \int_{\Xi} \psi(n,t) \ \mathcal{R}\varphi(n,t) \mathrm{d}n \mathrm{d}t = (-1)^m \langle \mathcal{R}\varphi, \psi \rangle.$$

Therefore,

$$\partial_t^m T_\varphi = (-1)^m \mathcal{R}\varphi \quad \text{in} \quad \mathcal{S}_0'(\Xi),$$

and, by (45),

$$T_\varphi = \mathcal{A}^m \partial_t^m T_\varphi = (-1)^m \mathcal{A}^m (\mathcal{R}\varphi) \quad \text{in} \quad \mathcal{S}_0'(\Xi).$$

Thus, there exists $p \in \mathcal{P}(\mathbb{R})$ such that

$$T_\varphi = (-1)^m \mathcal{A}^m (\mathcal{R}\varphi) + p \quad \text{in} \quad \mathcal{S}'(\Xi).$$

Hence,

$$T_\varphi(n,t) = (-1)^m \mathcal{A}^m(\mathcal{R}\varphi)(n,t) + p(t)$$

for almost every $(n,t) \in \Xi$, and therefore for every $(n,t) \in \Xi$ by continuity. We now show that the polynomial p has to vanish everywhere. Indeed, by the dominated convergence theorem,

$$\begin{aligned}\lim_{t\to+\infty} |T_\varphi(n,t)| &\leq \lim_{t\to+\infty} \int\limits_{\mathbb{R}^d} |\sigma_m(n\cdot x - t)||\varphi(x)|\mathrm{d}x \\ &= \lim_{t\to+\infty} \frac{1}{(m-1)!} \int\limits_{n\cdot x\geq t} (n\cdot x - t)^{m-1}|\varphi(x)|\mathrm{d}x \\ &\leq \lim_{t\to+\infty} \frac{1}{(m-1)!} \int\limits_{n\cdot x\geq t} |x|^{m-1}|\varphi(x)|\mathrm{d}x = 0.\end{aligned}$$

Furthermore, $t \mapsto \mathcal{A}^m(\mathcal{R}\varphi)(n,t) \in \mathcal{S}_0(\mathbb{R})$, and thus $\lim_{t\to+\infty} \mathcal{A}^m(\mathcal{R}\varphi)(n,t) = 0$. Hence, we can conclude that $p = 0$ and

$$T_\varphi(n,t) = (-1)^m \mathcal{A}^m(\mathcal{R}\varphi)(n,t)$$

for every $(n,t) \in \Xi$. Observing that

$${}_{\mathcal{S}_0'(\mathbb{R}^d)}\langle \rho_m(\cdot,n,t), \varphi\rangle_{\mathcal{S}_0(\mathbb{R}^d)} = \beta(n,t)T_\varphi(n,t),$$

the claim follows. □

The space $\mathcal{M}(\Xi)$ is clearly a subspace of $\mathcal{S}'(\Xi)$. The following simple lemma shows that it is a subspace of $\mathcal{S}_0'(\Xi)$.

Lemma 5.4. *Let $\mu, \mu' \in \mathcal{M}(\Xi)$ be such that $\mu = \mu'$ in $\mathcal{S}_0'(\Xi)$, then $\mu = \mu'$ in $\mathcal{M}(\Xi)$.*

Proof. Since $\mathcal{S}_0'(\Xi) \simeq \mathcal{S}'(\Xi)/\mathcal{P}(\mathbb{R})$ (see Appendix A), the equality $\mu = \mu'$ in $\mathcal{S}_0'(\Xi)$ means there exists a polynomial $p \in \mathcal{P}(\mathbb{R})$ such that $\mu' = \mu + p$ in $\mathcal{S}'(\Xi)$. But p must be 0 since μ, μ' are finite measures. Hence, $\mu' = \mu$ in $\mathcal{S}'(\Xi)$ and, a fortiori, in $\mathcal{M}(\Xi)$. □

The next result shows that $\|\cdot\|_{\mathrm{TV}}$ is invariant under symmetrization.

Lemma 5.5. *Let $\mu \in \mathcal{M}(\Xi)$. Then*

$$\|\mu^\vee\|_{\mathrm{TV}} = \|\mu\|_{\mathrm{TV}}.$$

Proof. Fix $\mu \in \mathcal{M}(\Xi)$. By definition of $\mu^\vee$ and $\psi^\vee$,

$$\int\limits_{\Xi} \psi(n,t)\ \mathrm{d}\mu^\vee(n,t) = \int\limits_{\Xi} \psi^\vee(n,t)\ \mathrm{d}\mu(n,t). \tag{48}$$

Indeed, using the above equality and $\|\psi^\vee\|_\infty = \|\psi\|_\infty$ for $\psi \in \mathrm{C}_0(\Xi)$, we have

$$\begin{aligned}\|\mu^\vee\|_{\mathrm{TV}} &= \sup\{\langle \mu^\vee, \psi\rangle \colon \psi \in \mathrm{C}_0(\Xi), \|\psi\|_\infty \le 1\}\\ &= \sup\{\langle \mu, \psi^\vee\rangle \colon \psi \in \mathrm{C}_0(\Xi), \|\psi\|_\infty \le 1\}\\ &= \sup\{\langle \mu, \psi\rangle \colon \psi \in \mathrm{C}_0(\Xi), \|\psi\|_\infty \le 1\} = \|\mu\|_{\mathrm{TV}}. \quad \square\end{aligned}$$

Equation (26b) shows that the functions $f \in \mathcal{B}_m$ are parametrized by the measures $\mu \in \mathcal{M}(\Xi)$. We now show that the even component of μ can be recovered by the Radon transform of f. The operator Λ^{d-1} is the Fourier multiplier defined by (62) and (66).

Lemma 5.6. *For every $f_\mu \in \mathcal{B}_m$,*

$$\frac{1}{2(2\pi)^{d-1}} \partial_t^m \Lambda^{d-1} \mathcal{R} f_\mu = \beta \ \frac{\mu + (-1)^m \mu^\vee}{2}, \tag{49}$$

where the equality holds in $\mathcal{S}_0'(\Xi)$.

Remark 5.7. Observe that $\Lambda^{d-1}\mathcal{R} f_\mu$ is an even distribution on Ξ. Furthermore, it is easy to check that

$$\partial_t^m \mathcal{S}_0'(\Xi)_{\mathrm{even}} \subseteq \begin{cases} \mathcal{S}_0'(\Xi)_{\mathrm{even}} & \text{if } m \text{ is even} \\ \mathcal{S}_0'(\Xi)_{\mathrm{odd}} & \text{if } m \text{ is odd} \end{cases}.$$

By (24b) β is even, so that $\beta\ (\mu + (-1)^m \mu^\vee)/2$ has the right parity. Without condition (24b), the statement of Lemma 5.6 holds true provided that the right hand side of (49) is replaced with $(\beta\mu + (-1)^m \beta^\vee \mu^\vee)/2$, which would make the decomposition of (35) more involved.

Proof. Assume first that m is even. As observed in Remark 5.7, both sides of (49) are even distributions. Thus, it is enough to check the equality on $\psi \in \mathcal{S}_0(\Xi)_{\mathrm{even}}$. We have

$$\begin{aligned} {}_{\mathcal{S}_0'(\Xi)}\langle \partial_t^m \Lambda^{d-1} \mathcal{R} f_\mu, \psi\rangle_{\mathcal{S}_0(\Xi)} &= (-1)^m {}_{\mathcal{S}_0'(\mathbb{R}^d)}\langle f_\mu, \mathcal{R}^* \Lambda^{d-1} \partial_t^m \psi\rangle_{\mathcal{S}_0(\mathbb{R}^d)} \\ &= (-1)^m \int\limits_{\mathbb{R}^d} f_\mu(x)\ \mathcal{R}^* \Lambda^{d-1} \partial_t^m \psi(x)\ \mathrm{d}x \\ &= (-1)^m \int\limits_{\mathbb{R}^d} \left(\int\limits_{\Xi} \rho_m(x, n, t)\ \mathrm{d}\mu(n, t) \right) \mathcal{R}^* \Lambda^{d-1} \partial_t^m \psi(x)\ \mathrm{d}x \\ &= (-1)^m \int\limits_{\Xi} \int\limits_{\mathbb{R}^d} \rho_m(x, n, t)\ \mathcal{R}^* \Lambda^{d-1} \partial_t^m \psi(x)\ \mathrm{d}x\ \mathrm{d}\mu(n, t) \\ &= (-1)^m \int\limits_{\Xi} \langle \rho_m(\cdot, n, t), \mathcal{R}^* \Lambda^{d-1} \partial_t^m \psi\rangle\ \mathrm{d}\mu(n, t). \end{aligned}$$

Proposition 5.3, the inversion formula (64) and (44) give that, for every $(n, t) \in \Xi$,

$$\begin{aligned} \langle \rho_m(\cdot, n, t), \mathcal{R}^* \Lambda^{d-1} \partial_t^m \psi\rangle &= (-1)^m \beta(n, t) \mathcal{A}^m \mathcal{R} \mathcal{R}^* \Lambda^{d-1} \partial_t^m \psi \\ &= (-1)^m 2(2\pi)^{d-1} \beta(n, t) \mathcal{A}^m \partial_t^m \psi \\ &= (-1)^m 2(2\pi)^{d-1} \beta(n, t) \psi(n, t). \end{aligned}$$

Thus, taking into account that both β (see (24b)) and ψ are even functions, we obtain

$$\begin{aligned}
{}_{\mathcal{S}_0'(\Xi)}\langle \partial_t^m \Lambda^{d-1}\mathcal{R} f_\mu, \psi\rangle_{\mathcal{S}_0(\Xi)} &= 2(2\pi)^{d-1}\int\limits_{\Xi} \beta(n,t)\psi(n,t)\ \mathrm{d}\mu(n,t) \\
&= 2(2\pi)^{d-1}\int\limits_{\Xi} \beta(n,t)\psi(n,t)\ \mathrm{d}\mu_{\mathrm{even}}(n,t) \\
&= 2(2\pi)^{d-1}{}_{\mathcal{S}_0'(\Xi)}\langle \beta\ \mu_{\mathrm{even}}, \psi\rangle_{\mathcal{S}_0(\Xi)},
\end{aligned}$$

which proves (49) for even m. If m is odd, the proof is very similar, observing that both sides of (49) are odd distributions, and thus checking the equality on $\psi \in \mathcal{S}_0(\Xi)_{\mathrm{odd}}$. Furthermore, $\partial_t^m\psi$ is an even function, so that $\partial_t^m\psi \in \mathcal{S}_0(\mathbb{P}^d)$, and $\beta\psi$ is an odd function, so that

$$\int\limits_{\Xi} \beta(n,t)\psi(n,t)\ \mathrm{d}\mu(n,t) = \int\limits_{\Xi} \beta(n,t)\psi(n,t)\ \mathrm{d}\mu_{\mathrm{odd}}(n,t). \qquad \square$$

The map $\mu \mapsto f_\mu$ is not injective and next result characterizes its kernel.

Lemma 5.8. *Let $\mu \in \mathcal{M}(\Xi)$. Then:*

(i) *if $f_\mu = 0$, then*

$$\mu^\vee = (-1)^{m+1}\mu \qquad \Longleftrightarrow \qquad \mu \in \begin{cases} \mathcal{S}_0'(\Xi)_{\mathrm{odd}} & \text{if } m \text{ is even} \\ \mathcal{S}_0'(\Xi)_{\mathrm{even}} & \text{if } m \text{ is odd} \end{cases};$$

(ii) *if $\mu^\vee = (-1)^{m+1}\mu$, then f_μ is a polynomial of degree less than m.*

Furthermore,

$$\mathcal{P}_m = \{p : \mathbb{R}^d \to \mathbb{R} : p \text{ is a polynomial of degree at most } m-1\},$$

where $\mathcal{P}_m$ is the space defined in Theorem 4.1.

Proof. Let $\tau = (\mu + (-1)^m\mu^\vee)/2$. If $f_\mu = 0$, then (49) implies that $\beta\tau = 0$ in $\mathcal{S}_0'(\Xi)$ and, by (24a), $\tau = 0$ in $\mathcal{S}_0'(\Xi)$ and, by Lemma 5.4, $\tau = 0$ in $\mathcal{M}(\Xi)$.

Assume that $\tau = 0$. Then (49) gives that

$$\partial_t^m \Lambda^{d-1}\mathcal{R} f_\mu = 0$$

in $\mathcal{S}_0'(\Xi)$. Equation (45) implies that ∂_t^m is injective, so that $\Lambda^{d-1}\mathcal{R} f_\mu = 0$ in $\mathcal{S}_0'(\Xi)$. By construction $\Lambda^{d-1}\mathcal{R} f_\mu \in \mathcal{S}_0'(\Xi)_{\mathrm{even}}$. Then, by Corollary A.11, we have that

$$f_\mu = \frac{1}{2(2\pi)^{d-1}}\mathcal{R}^*\Lambda^{d-1}\mathcal{R} f_\mu = 0 \quad \text{in} \quad \mathcal{S}_0'(\mathbb{R}^d),$$

or equivalently, there exists $p \in \mathcal{P}(\mathbb{R})$ such that $f_\mu = p$ in $\mathcal{S}'(\mathbb{R}^d)$. Hence,

$$f_\mu(x) = p(x)$$

for almost every $x \in \mathbb{R}^d$, and thus for every $x \in \mathbb{R}^d$ by continuity. But since the elements of $\mathcal{B}_m$ are functions of at most $m-1$ polynomial growth (see (43)), we obtain that f_μ is a polynomial of degree less than m. We now prove the last claim.

By item (ii), $\mathcal{P}_m$ is a subspace of the finite-dimensional vector space of polynomials of degree smaller than m. Now, let $\nu = (\delta_{(n,t)} + (-1)^{m+1}\delta_{(-n,-t)})/2$ with $(n,t) \in \Xi$. Then, by (26b) and (24b),

$$f_\nu(x) = \int_\Xi \sigma_m(n' \cdot x - t')\beta(n',t')\, \mathrm{d}\nu(n',t') = \beta(n,t)\frac{(n\cdot x - t)^{m-1}}{2(m-1)!},$$

where in the last equality we used

$$\max\{0,t\}^{m-1} + (-1)^{m+1}\max\{0,-t\}^{m-1} = t^{m-1}.$$

Then

$$\operatorname{span}\{(n\cdot x - t)^{m-1} \colon (n,t)\in\Xi\} \subseteq \mathcal{P}_m.$$

However, it is known that the left hand side of the above inequality is the space of polynomials of degree less or equal $m-1$, so that the claim is proved. □

We are now ready to prove Theorem 4.1 and Corollary 4.2.

Proof of Theorem 4.1. We prove the statements for an even m (if m is odd the proof is similar). We regard $\mathcal{Q}_m$ and $\mathcal{P}_m$ as reproducing kernel Banach spaces with the norms

$$\|f\|_{\mathcal{Q}_m} = \inf\{\|\mu\|_{\mathrm{TV}} : \mu \in \mathcal{M}(\Xi), \mu^\vee = (-1)^m\mu, f = f_\mu\}, \tag{50a}$$

$$\|f\|_{\mathcal{P}_m} = \inf\{\|\mu\|_{\mathrm{TV}} : \mu \in \mathcal{M}(\Xi), \mu^\vee = (-1)^{m+1}\mu, f = f_\mu\}. \tag{50b}$$

Note that in principle these norms induce respectively on $\mathcal{Q}_m$ and $\mathcal{P}_m$ a finer topology than the one induced by the norm $\|\cdot\|_{\mathcal{B}_m}$. Fix $f \in \mathcal{B}_m$. By (26a), there exists $\mu \in \mathcal{M}(\Xi)$ such that $f = f_\mu$. Define

$$\tau = \frac{\mu + \mu^\vee}{2} \in \mathcal{M}(\Xi)_{\mathrm{even}}, \qquad \nu = \frac{\mu - \mu^\vee}{2} \in \mathcal{M}(\Xi)_{\mathrm{odd}},$$

and compare with (28) taking into account that m is even. By linearity of the representation (26b),

$$f = f_\tau + f_\nu,$$

whereas item (i) of Lemma 5.8 gives

$$\mathcal{Q}_m \cap \mathcal{P}_m = \{0\}, \tag{51}$$

so that

$$\mathcal{B}_m = \mathcal{Q}_m + \mathcal{P}_m,$$

and

$$f_\tau = P_{\mathcal{Q}_m} f, \qquad f_\nu = P_{\mathcal{P}_m} f, \tag{52}$$

which shows item (ii). The fact that $\mathcal{P}_m$ is the space of polynomials of degree less or equal $m-1$ is the content of item (ii) of Lemma 5.8, whereas item (i) is the content of item (iii) of Lemma 5.1. Since τ is the even part of μ, (49) gives

$$\frac{1}{2(2\pi)^{d-1}\beta}\partial_t^m \Lambda^{d-1}\mathcal{R}f = \frac{\mu+\mu^\vee}{2} = \tau,$$

hence (29) holds true.

If $f = f_{\mu'}$ for another $\mu' \in \mathcal{M}(\Xi)$, by Lemma 5.8 we have

$$\mu' = \tau + \nu', \qquad \tau = \frac{\mu' + (\mu')^\vee}{2}, \qquad f_{\nu'} = f_\nu,$$

for some odd measure ν'. Taking into account the above equalities, (26c) gives

$$\begin{aligned} \|f\|_{\mathcal{B}_m} &= \inf\{\|\tau+\nu'\|_{\mathrm{TV}} : \nu' \in \mathcal{M}(\Xi)_{\mathrm{odd}}, f_{\nu'} = f_\nu\} \\ &\leq \inf\{\|\tau\|_{\mathrm{TV}} + \|\nu'\|_{\mathrm{TV}} \colon \nu' \in \mathcal{M}(\Xi)_{\mathrm{odd}}, f_{\nu'} = f_\nu\} \\ &= \|\tau\|_{\mathrm{TV}} + \inf\{\|\nu'\|_{\mathrm{TV}} \colon \nu' \in \mathcal{M}(\Xi)_{\mathrm{odd}}, f_{\nu'} = f_\nu\} \\ &= \|f_\tau\|_{\mathcal{Q}_m} + \|f_\nu\|_{\mathcal{P}_m}, \end{aligned} \tag{53}$$

where the second inequality is a consequence of the triangular inequality, the third one is due to the fact that τ is even and ν' is odd, and the last equality is a consequence of (50a) and (50b) observing that τ is the unique even measure such that $f_\tau = P_{\mathcal{Q}_m} f$, so that

$$\|f_\tau\|_{\mathcal{Q}_m} = \|\tau\|_{\mathrm{TV}}. \tag{54}$$

Furthermore, by Lemma 5.5 we have that

$$\|f_\tau\|_{\mathcal{Q}_m} \leq \|\frac{\mu' + (\mu')^\vee}{2}\|_{\mathrm{TV}} \leq \|\mu'\|_{\mathrm{TV}}, \qquad \|f_\nu\|_{\mathcal{P}_m} \leq \|\frac{\mu' - (\mu')^\vee}{2}\|_{\mathrm{TV}} \leq \|\mu'\|_{\mathrm{TV}}.$$

Therefore, taking the infimum over all measures μ' such that $f_{\mu'} = f$, we get

$$\|f_\tau\|_{\mathcal{Q}_m} \leq \|f\|_{\mathcal{B}_m}, \qquad \|f_\nu\|_{\mathcal{P}_m} \leq \|f\|_{\mathcal{B}_m}, \tag{55}$$

which, together with (53), gives

$$\|f\|_{\mathcal{B}_m} \leq \|f_\tau\|_{\mathcal{Q}_m} + \|f_\nu\|_{\mathcal{P}_m} \leq 2\|f\|_{\mathcal{B}_m}. \tag{56}$$

If $f \in \mathcal{Q}_m$, then $f = f_\tau$ and by equations (56) and (55) we have that

$$\|f\|_{\mathcal{B}_m} \leq \|f\|_{\mathcal{Q}_m} \leq \|f\|_{\mathcal{B}_m}.$$

So that, by (54)

$$\|f\|_{\mathcal{B}_m} = \|f\|_{\mathcal{Q}_m} = \|\tau\|_{\mathrm{TV}},$$

which proves (31). If $f \in \mathcal{P}_m$, then $\tau = 0$ and, as above,

$$\|f\|_{\mathcal{B}_m} = \|f\|_{\mathcal{P}_m} = \inf\{\|\nu\|_{\mathrm{TV}} \colon \nu \in \mathcal{M}(\Xi)_{\mathrm{odd}}, f_\nu = f\},$$

which is (32). Finally, (31) and (32) together with (56) give equation (30). This also implies that $\mathcal{Q}_m$ and $\mathcal{P}_m$ are closed subspaces of $\mathcal{B}_m$.

We finally prove item (iv). Fix a distribution T as in the statement. By assumption (33) and Lemma 5.4, there exists a unique even measure τ such that

$$\tau = \frac{1}{2(2\pi)^{d-1}\beta} \partial_t^m \Lambda^{d-1} \mathcal{R} T,$$

hence $f_\tau \in \mathcal{Q}_m$. Equation (34) ensures that there exists $\nu \in \mathcal{M}(\Xi)_{\text{odd}}$ such that $T - f_\tau = f_\nu$. Setting $\mu = \tau + \nu$, we get

$$T - f_\mu = (T - f_\tau) - f_\nu = 0,$$

which proves (iv). □

Proof of Corollary 4.2. Reasoning as in the last part of the previous proof, and again assuming that m is even, (49) implies that

$$\partial_t^m \Lambda^{d-1} \mathcal{R}(T - f_\tau) = 0$$

in $\mathcal{S}_0'(\Xi)_{\text{even}}$. The injectivity of the operator $\partial_t^m \Lambda^{d-1} \mathcal{R}$ gives that $(T - f_\tau) = 0$ in $\mathcal{S}_0'(\mathbb{R}^d)$, *i.e.* there exists a polynomial p such that $T - f_\tau = p$ in $\mathcal{S}'(\mathbb{R}^d)$. □

Acknowledgments

The authors would like to thank Filippo De Mari, and especially Jaouad Mourtada and Koen Sanders for useful discussions. Indeed, the ideas in this paper were initially explored as Koen's summer project under the co-supervision of Jaouad. L.R. acknowledges support from the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. L.R. also acknowledges the financial support of the European Research Council (grant SLING 819789), the AFOSR projects FA9550-18-1-7009, FA9550-17-1-0390 and BAA-AFRL-AFOSR-2016-0007 (European Office of Aerospace Research and Development), and the EU H2020-MSCA-RISE project NoMADS - DLV-777826. E.D.V. is a member of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM).

Appendix A. Radon transform: review

We start recalling the function spaces that come into play. Let $d \in \mathbb{N}$, $d \geq 1$. We use the notation $\langle x \rangle = (1 + |x|^2)^{\frac{1}{2}}$. We denote by $\mathcal{S}(\mathbb{R}^d)$ the Schwartz space of rapidly decreasing functions. We recall that a function $\varphi \colon \mathbb{R}^d \to \mathbb{C}$ belongs to $\mathcal{S}(\mathbb{R}^d)$ if $\varphi \in C^\infty(\mathbb{R}^d)$ and

$$\rho_{m,\alpha}(\varphi) = \sup_{x \in \mathbb{R}^d} \langle x \rangle^m |\partial^\alpha \varphi(x)| < +\infty, \qquad \forall m, \alpha \in \mathbb{N}^d. \tag{57}$$

We endow $\mathcal{S}(\mathbb{R}^d)$ with the topology induced by the family of seminorms $\{\rho_{m,\alpha}\}_{m,\alpha \in \mathbb{N}^d}$, which makes $\mathcal{S}(\mathbb{R}^d)$ a Fréchet space. Its dual space $\mathcal{S}'(\mathbb{R}^d)$ is known as the space of tempered distributions. We use the notation $\mathcal{P}(\mathbb{R}^d)$ for the space of all polynomials on $\mathbb{R}^d$ and we denote by $\mathcal{S}_0(\mathbb{R}^d)$ the space of functions in $\mathcal{S}(\mathbb{R}^d)$ that are orthogonal to all polynomials, *i.e.*

$$\mathcal{S}_0(\mathbb{R}^d) = \left\{ \varphi \in \mathcal{S}(\mathbb{R}^d) : \int\limits_{\mathbb{R}^d} \varphi(x) p(x) \mathrm{d}x = 0, \ \forall p \in \mathcal{P}(\mathbb{R}^d) \right\}. \tag{58}$$

The space $\mathcal{S}_0(\mathbb{R}^d)$ is called the Lizorkin test function space. It is a closed subspace of $\mathcal{S}(\mathbb{R}^d)$ and we endow it with the relative topology inherited from $\mathcal{S}(\mathbb{R}^d)$. Its dual space $\mathcal{S}_0'(\mathbb{R}^d)$ of Lizorkin distributions is topologically isomorphic to the quotient space $\mathcal{S}'(\mathbb{R}^d)/\mathcal{P}(\mathbb{R}^d)$, see *e.g.* [23, Chapter 1, Section 25].

Lemma A.1 *([23, Lemma 6.0.4]). Let $\varphi \in \mathcal{S}(\mathbb{R})$. Then $\varphi \in \mathcal{S}_0(\mathbb{R})$ if and only if, for every $k \in \mathbb{N}$,*

$$\lim_{\omega \to 0} \frac{\mathcal{F}\varphi(\omega)}{|\omega|^k} = 0.$$

As a consequence of Lemma A.1, the Fourier transform maps $\mathcal{S}_0(\mathbb{R})$ into the space $\hat{\mathcal{S}}_0(\mathbb{R})$ of rapidly decreasing functions that vanish in zero together with all of their partial derivatives, *i.e.*

$$\hat{\mathcal{S}}_0(\mathbb{R}) = \{\varphi \in \mathcal{S}(\mathbb{R}) : \partial^m \varphi(0) = 0, \ \forall m \in \mathbb{N}\}.$$

Recall that $\Xi = S^{d-1} \times \mathbb{R}$. In analogy with $\mathcal{S}(\mathbb{R}^d)$, we denote by $\mathcal{S}(\Xi)$ the space of functions in $C^\infty(\Xi)$ such that

$$\rho_{k,l,D}(\psi) = \sup_{n \in S^{d-1}, t \in \mathbb{R}} \langle t \rangle^k \left| \frac{\mathrm{d}^l}{\mathrm{d}t^l} D\psi(n,t) \right| < +\infty,$$

for every $k, l \in \mathbb{N}$ and for every differentiable operator D on S^{d-1}. We endow $\mathcal{S}(\Xi)$ with the topology induced by the family of seminorms $\rho_{k,l,D}$, and we denote by $\mathcal{S}'(\Xi)$ its topological dual space. In analogy with the Lizorkin test function space, $\mathcal{S}_0(\Xi)$ denotes the set of functions $\psi \in \mathcal{S}(\Xi)$ such that

$$\int_{\mathbb{R}} \psi(n,t) p(t) \mathrm{d}t = 0, \qquad \forall p \in \mathcal{P}(\mathbb{R}), n \in S^{d-1}. \tag{59}$$

Note that the integrals in (59) are finite since the functions $t \mapsto t^k \psi(n,t)$ belong to $L^1(\mathbb{R})$ for every $k \in \mathbb{N}$ and $n \in S^{d-1}$. Then, by Lemma A.1, condition (59) is equivalent to requiring that

$$\lim_{\omega \to 0} \frac{\mathcal{F}\psi(n,\omega)}{|\omega|^k} = 0, \qquad \forall k \in \mathbb{N}, n \in S^{d-1},$$

where $\mathcal{F}$ denotes the Fourier transform acting on the second variable. We further refer to [21] for a complete exposition of the function spaces introduced above.

Remark A.2. Usually, the Radon transform $\mathcal{R}f$ of a function $f : \mathbb{R}^d \to \mathbb{C}$ is defined on the space $\mathbb{P}^d$ of all hyperplanes in $\mathbb{R}^d$. As seen in Section 4.2.1, Ξ is the double covering of $\mathbb{P}^d$ with respect to the equivalence relation (39). Hence, we can identify functions and distributions on $\mathbb{P}^d$ with even functions and even distributions on Ξ and we can define the distribution Radon transform as a map from $\mathcal{S}_0'(\mathbb{R}^d)$ onto $\mathcal{S}_0'(\Xi)_{\text{even}} \simeq \mathcal{S}_0'(\mathbb{P}^d)$ and its dual $\mathcal{R}^*$ as a map from $\mathcal{S}_0'(\Xi)_{\text{even}} \simeq \mathcal{S}_0'(\mathbb{P}^d)$ into $\mathcal{S}_0'(\mathbb{R}^d)$. We adopt this setting since the space $\mathcal{S}_0'(\mathbb{P}^d)$ is replaced by $\mathcal{S}_0'(\Xi)_{\text{odd}}$ to deal with odd m, see Theorem 4.1.

We briefly recall the notion of even and odd distributions. For all functions $\psi : \Xi \to \mathbb{C}$, we set

$$\psi^\vee : \Xi \to \mathbb{C}, \qquad \psi^\vee(n,t) = \psi(-n,-t), \qquad (n,t) \in \Xi.$$

It is easy to check that

$$\mathcal{S}_0(\Xi) \ni \psi \mapsto \psi^\vee \in \mathcal{S}_0(\Xi)$$

is a well-defined continuous involution and, by duality, it defines an involution on $\mathcal{S}_0'(\Xi)$

$$\mathcal{S}_0'(\Xi) \ni g \mapsto g^\vee \in \mathcal{S}_0'(\Xi).$$

We set

$$\mathcal{S}_0(\Xi)_{\text{even}} = \{\psi \in \mathcal{S}_0(\Xi) : \psi^\vee = \psi\}, \qquad \mathcal{S}_0(\Xi)_{\text{odd}} = \{\psi \in \mathcal{S}_0(\Xi) : \psi^\vee = -\psi\},$$
$$\mathcal{S}_0'(\Xi)_{\text{even}} = \{g \in \mathcal{S}_0'(\Xi) : g^\vee = g\}, \qquad \mathcal{S}_0'(\Xi)_{\text{odd}} = \{g \in \mathcal{S}_0'(\Xi) : g^\vee = -g\},$$

which are closed subsets of $\mathcal{S}_0(\Xi)$ and $\mathcal{S}_0'(\Xi)$, respectively. Moreover,

$$\mathcal{S}_0(\Xi) = \mathcal{S}_0(\Xi)_{\text{even}} + \mathcal{S}_0(\Xi)_{\text{odd}}, \qquad \mathcal{S}_0(\Xi)_{\text{even}} \cap \mathcal{S}_0(\Xi)_{\text{odd}} = \{0\},$$
$$\mathcal{S}_0'(\Xi) = \mathcal{S}_0'(\Xi)_{\text{even}} + \mathcal{S}_0'(\Xi)_{\text{odd}}, \qquad \mathcal{S}_0'(\Xi)_{\text{even}} \cap \mathcal{S}_0'(\Xi)_{\text{odd}} = \{0\},$$

where the maps

$$\mathcal{S}_0(\Xi)_{\text{even}} \times \mathcal{S}_0(\Xi)_{\text{odd}} \ni (\psi_{\text{even}}, \psi_{\text{odd}}) \mapsto \psi_{\text{even}} + \psi_{\text{odd}} \in \mathcal{S}_0(\Xi),$$
$$\mathcal{S}_0'(\Xi)_{\text{even}} \times \mathcal{S}_0'(\Xi)_{\text{odd}} \ni (g_{\text{even}}, g_{\text{odd}}) \mapsto g_{\text{even}} + g_{\text{odd}} \in \mathcal{S}_0'(\Xi)$$

are topological isomorphisms. A simple calculation shows that

$$\mathcal{S}_0'(\Xi)_{\text{even}} \simeq (\mathcal{S}_0(\Xi)_{\text{even}})',$$
$$\mathcal{S}_0(\Xi)'_{\text{odd}} \simeq (\mathcal{S}_0(\Xi)_{\text{odd}})',$$

which implies that $\mathcal{S}_0'(\Xi)_{\text{even}} \simeq \mathcal{S}_0'(\mathbb{P}^d)$ under the identification $\mathcal{S}_0(\Xi)_{\text{even}} = \mathcal{S}_0(\mathbb{P}^d)$, as claimed in Remark A.2. With this setting, we can recall the definition of the Radon transform and its dual.

Definition A.3. The Radon transform of $\varphi \in L^1(\mathbb{R}^d)$ is the function $\mathcal{R}\varphi \colon \Xi \to \mathbb{C}$ defined by

$$\mathcal{R}\varphi(n,t) = \int\limits_{n\cdot x = t} \varphi(x)\mathrm{d}m(x), \qquad \text{for a.e. } (n,t) \in \Xi,$$

where m is the Euclidean measure on the hyperplane with equation $n \cdot x = t$.

Since the pairs (n,t) and $(-n,-t)$ define the same hyperplane, clearly the Radon transform is an even function, *i.e.*

$$(\mathcal{R}\varphi)^\vee = \mathcal{R}\varphi. \tag{60}$$

Theorem A.4 *([20, Corollary 4.2]). The Radon transform is a continuous injective operator from $\mathcal{S}_0(\mathbb{R}^d)$ onto $\mathcal{S}_0(\Xi)_{\text{even}}$.*

We now introduce the dual Radon transform, also known as back-projection. While the Radon transform is defined for any pair (n,t) as the integral over the set of points belonging to the hyperplane with equation $n \cdot x = t$, the dual Radon transform is defined for any given point $x \in \mathbb{R}^d$ as the integral over the set of hyperplanes passing through x, which corresponds to the set of pairs $\{(n, n\cdot x) : n \in S^{d-1}\} \subseteq \Xi$.

Definition A.5. The dual Radon transform (or back-projection) of $\psi \in L^\infty(\Xi)$ is the L^∞ function $\mathcal{R}^*\psi : \mathbb{R}^d \to \mathbb{C}$ defined by

$$\mathcal{R}^*\psi(x) = \int\limits_{S^{d-1}} \psi(n, n\cdot x)\mathrm{d}n, \qquad x \in \mathbb{R}^d,$$

where $\mathrm{d}n$ is the spherical measure on S^{d-1}.

Note that, if ψ is an odd function, clearly $\mathcal{R}^*\psi = 0$ since dn is invariant under reflection.

Theorem A.6 *([20, Corollary 4.2]). The dual Radon transform is a continuous injective operator from* $\mathcal{S}_0(\Xi)_{\text{even}}$ *onto* $\mathcal{S}_0(\mathbb{R}^d)$.

We refer to [27, Corollary 6.1] for an alternative proof of the continuity of the operators $\mathcal{R}\colon \mathcal{S}_0(\mathbb{R}^d) \to \mathcal{S}_0(\Xi)_{\text{even}}$ and $\mathcal{R}^*\colon \mathcal{S}_0(\Xi)_{\text{even}} \to \mathcal{S}_0(\mathbb{R}^d)$ based on the relation existing between Radon, ridgelet and wavelet transforms.

Proposition A.7 *([33, Chapter II]). For every* $\varphi \in L^1(\mathbb{R}^d)$ *and* $\psi \in L^\infty(\Xi)$,

$$\int\limits_{\mathbb{R}^d} \varphi(x)\mathcal{R}^*\psi(x)\mathrm{d}x = \int\limits_{\Xi} \mathcal{R}\varphi(n,t)\psi(n,t)\mathrm{d}n\mathrm{d}t. \tag{61}$$

The duality relation (61) can be exploited to extend $\mathcal{R}$ and $\mathcal{R}^*$ on distribution spaces [21,22,27].

Definition A.8. The Radon transform of $f \in \mathcal{S}_0'(\mathbb{R}^d)$ is the continuous linear functional $\mathcal{R}f$ on $\mathcal{S}_0(\Xi)_{\text{even}}$ defined by

$$\langle \mathcal{R}f, \psi\rangle = \langle f, \mathcal{R}^*\psi\rangle, \qquad \psi \in \mathcal{S}_0(\Xi)_{\text{even}}.$$

Analogously, the dual Radon transform of $g \in \mathcal{S}_0'(\Xi)_{\text{even}}$ is the continuous linear functional on $\mathcal{S}_0(\mathbb{R}^d)$ defined by

$$\langle \mathcal{R}^*g, \varphi\rangle = \langle g, \mathcal{R}\varphi\rangle, \qquad \varphi \in \mathcal{S}_0(\mathbb{R}^d).$$

Note that $\mathcal{R} : \mathcal{S}_0'(\mathbb{R}^d) \to \mathcal{S}_0'(\Xi)_{\text{even}}$ and $\mathcal{R}^* : \mathcal{S}_0'(\Xi)_{\text{even}} \to \mathcal{S}_0'(\mathbb{R}^d)$ are well defined and weakly continuous thanks to Theorem A.6 and Theorem A.4, respectively. We next recall the most commonly used inversion formula for the Radon transform, known as Filtered Back Projection. To state the formula, we first need to introduce the positive symmetric operator $\Lambda^{d-1} : \mathcal{S}(\Xi) \to C^\infty(\Xi)$ defined by

$$\Lambda^{d-1}\psi(n,t) = \begin{cases} (-1)^{\frac{d-1}{2}}\partial_t^{d-1}\psi(n,t) & d \text{ odd} \\ (-1)^{\frac{d-2}{2}}\mathscr{H}\partial_t^{d-1}\psi(n,t) & d \text{ even} \end{cases}, \tag{62}$$

where the Hilbert transform $\mathscr{H}$ acts only on the second variable. The operator Λ^{d-1} is also known as ramp filter.

Theorem A.9 *([21, Chapter I, Theorems 3.6 and 3.5]). For every* $\varphi \in \mathcal{S}(\mathbb{R}^d)$,

$$\varphi = \frac{1}{2(2\pi)^{d-1}}\mathcal{R}^*\Lambda^{d-1}\mathcal{R}\varphi. \tag{63}$$

For every $g \in \mathcal{S}_0(\Xi)_{\text{even}}$,

$$g = \frac{1}{2(2\pi)^{d-1}}\Lambda^{d-1}\mathcal{R}\mathcal{R}^*g. \tag{64}$$

In [22, Proposition 4.3], the inversion formula (63) has been extended to the space $\mathcal{D}_{L^1}'(\mathbb{R}^d)$ of Schwartz integrable distributions [48], which embeds densely in $\mathcal{S}_0'(\mathbb{R}^d)$. In Sections 4 and 5, we largely exploit the extensions of the inversion formulae (63) and (64) to the larger spaces of Lizorkin distributions $\mathcal{S}_0'(\mathbb{R}^d)$ and

$\mathcal{S}_0'(\Xi)_{\text{even}}$, respectively (cf. Corollary A.11). It is worth observing that the Hilbert transform appears in the expression of the operator Λ^{d-1} only when the dimension d is even. This difference is crucial in the Radon transform theory. For odd dimension d, Λ^{d-1} is a differential operator and it is therefore clear that it maps $\mathcal{S}(\Xi)$ continuously into itself. This no longer holds if d is even, because the Hilbert transform maps $\mathcal{S}(\mathbb{R})$ into $C^\infty(\mathbb{R})$, but not into $\mathcal{S}(\mathbb{R})$ [30]. A more satisfactory situation is obtained if we restrict our attention to the smaller space of functions $\mathcal{S}_0(\Xi)$. The continuity of fractional derivatives from $\mathcal{S}_0(\mathbb{R})$ into itself is a standard result in harmonic analysis [24, Chapter 2,§ 8]. The same property applies to the Hilbert transform and we report the proof for completeness.

Lemma A.10. *The Hilbert transform maps $\mathcal{S}_0(\mathbb{R})$ continuously into itself, and therefore Λ^{d-1} maps $\mathcal{S}_0(\Xi)_{\text{even}}$ continuously into itself for every $d \geq 1$.*

Proof. We start showing that $\mathscr{H}$ maps $\mathcal{S}_0(\mathbb{R})$ into $\mathcal{S}(\mathbb{R})$. Let $\varphi \in \mathcal{S}_0(\mathbb{R})$. We already know that $\mathscr{H}\varphi \in C^\infty(\mathbb{R})$. Thus, it remains to show that $\mathscr{H}\varphi$ is a rapidly decreasing function, or equivalently that $\mathcal{F}[\mathscr{H}\varphi] \in \mathcal{S}(\mathbb{R})$. We recall that $\mathscr{H}$ maps $\mathcal{S}(\mathbb{R})$ into $L^2(\mathbb{R})$, and for every $\varphi \in \mathcal{S}(\mathbb{R})$ it satisfies

$$\mathcal{F}[\mathscr{H}\varphi](\omega) = -i\,\mathrm{sgn}(\omega)\mathcal{F}\varphi(\omega), \quad \text{for a.e. } \omega \in \mathbb{R}.$$

Hence, we have that $\mathcal{F}[\mathscr{H}\varphi] \in C^\infty(\mathbb{R} \setminus \{0\})$, and for every $l \in \mathbb{N}$

$$\partial_\omega^l \mathcal{F}[\mathscr{H}\varphi](\omega) = -i\,\mathrm{sgn}(\omega)\partial_\omega^l \mathcal{F}\varphi(\omega), \qquad \omega \neq 0. \tag{65}$$

Since $\varphi \in \mathcal{S}_0(\mathbb{R})$, $\partial_\omega^l \mathcal{F}\varphi(0) = 0$ for every $l \in \mathbb{N}$, and $\mathcal{F}[\mathscr{H}\varphi]$ can be extended together with all its derivatives to continuous functions on $\mathbb{R}$. Therefore, $\mathcal{F}[\mathscr{H}\varphi] \in C^\infty(\mathbb{R})$ and hence $\mathscr{H}\varphi \in \mathcal{S}(\mathbb{R})$. In fact, $\mathscr{H}\varphi \in \mathcal{S}_0(\mathbb{R})$. Indeed, since $\varphi \in \mathcal{S}_0(\mathbb{R})$, for every $k \in \mathbb{N}$

$$\lim_{\omega \to 0} \frac{\mathcal{F}[\mathscr{H}\varphi](\omega)}{\omega^k} = \lim_{\omega \to 0} \frac{-i\ \mathrm{sgn}(\omega)\mathcal{F}\varphi(\omega)}{\omega^k} = 0,$$

which implies $\mathscr{H}\varphi \in \mathcal{S}_0(\mathbb{R})$ by Lemma A.1. We now show that $\mathscr{H}$ is continuous from $\mathcal{S}_0(\mathbb{R})$ into itself. In view of (65), for every $\omega \in \mathbb{R}$ and $m, \alpha \in \mathbb{N}$ we have

$$\langle \omega \rangle^m |\partial_\omega^\alpha \mathcal{F}\mathscr{H}\varphi(\omega)| = \langle \omega \rangle^m |\partial_\omega^\alpha \mathcal{F}\varphi(\omega)|.$$

The claim follows by observing that $\rho_{m,\alpha}(\mathcal{F}\varphi)$, $m, \alpha \in \mathbb{N}$, defines a basis of seminorms for the topology of $\mathcal{S}_0(\mathbb{R})$. Therefore, since $\mathcal{S}_0(\mathbb{R})$ is closed under differentiation and since $\mathscr{H}$ maps $\mathcal{S}_0(\mathbb{R})$ continuously into itself, it is clear from the definition that Λ^{d-1} maps $\mathcal{S}_0(\Xi)$ continuously into itself for every $d \geq 1$. Furthermore, if $g \in \mathcal{S}(\Xi)_{\text{even}}$, then $\Lambda^{d-1}g$ satisfies the symmetry condition (60) [21, Chapter I, Section 3]. Therefore, Λ^{d-1} maps $\mathcal{S}_0(\Xi)_{\text{even}}$ into itself for every $d \geq 1$. □

Thanks to Lemma A.10, we can define the weakly continuous operator $\Lambda^{d-1}\colon \mathcal{S}_0'(\Xi)_{\text{even}} \to \mathcal{S}_0'(\Xi)_{\text{even}}$ given by

$$\langle \Lambda^{d-1}g, \varphi \rangle = \langle g, \Lambda^{d-1}\varphi \rangle, \qquad g \in \mathcal{S}_0'(\Xi)_{\text{even}}, \varphi \in \mathcal{S}_0(\Xi)_{\text{even}}. \tag{66}$$

Definition A.8 together with equation (66) yields the following extension of the inversion formulas (63) and (64).

Corollary A.11. *For every $f \in \mathcal{S}_0'(\mathbb{R}^d)$,*

$$f = \frac{1}{2(2\pi)^{d-1}} \mathcal{R}^* \Lambda^{d-1} \mathcal{R} f.$$

For every $g \in \mathcal{S}_0'(\Xi)_{\mathrm{even}}$*,*

$$g = \frac{1}{2(2\pi)^{d-1}} \Lambda^{d-1} \mathcal{R}\mathcal{R}^* g.$$

Appendix B. Sparse solutions in variational problems

In this section we collect some results from [8] that we use in our paper. We start recalling the definition of extremal point.

Definition B.1. Let Q be a convex subset of a locally convex space. A point $q \in Q$ is called extremal if $Q \setminus \{q\}$ is convex. We denote the set of extremal points of Q by $\mathrm{Ext}(Q)$.

While extremal points are difficult to characterize in general, the following result is fairly standard (see [8, Proposition 4.1]). We report the proof for the reader's convenience.

Lemma B.2. *Let* Θ *be a (Hausdorff) locally compact second countable topological space, and let*

$$B = \{\mu \in \mathcal{M}(\Theta) : \|\mu\|_{\mathrm{TV}} \leq 1\}$$

be the unit ball in $\mathcal{M}(\Theta)$ *associated with the total variation norm. Then*

$$\mathrm{Ext}(B) = \{\pm\delta_\theta : \theta \in \Theta\}.$$

Proof. We start showing that $\{\pm\delta_\theta : \theta \in \Theta\} \subseteq \mathrm{Ext}(B)$. Let $\theta \in \Theta$ and $\alpha \in \{-1, 1\}$. We suppose that there exist $t \in (0, 1)$, $\mu_1, \mu_2 \in B$ such that

$$\alpha\delta_\theta = t\mu_1 + (1-t)\mu_2, \tag{67}$$

and we want to show that necessarily $\alpha\delta_\theta = \mu_1 = \mu_2$. We observe that the total variation measures $|\mu_1|$, $|\mu_2|$ are probability measures. Indeed, if we suppose on the contrary that $\|\mu_1\|_{\mathrm{TV}}, \|\mu_2\|_{\mathrm{TV}} < 1$, then

$$\|\alpha\delta_\theta\|_{\mathrm{TV}} \leq t\|\mu_1\|_{\mathrm{TV}} + (1-t)\|\mu_2\|_{\mathrm{TV}} < 1,$$

which yields a contradiction. Furthermore,

$$\delta_\theta = t|\mu_1| + (1-t)|\mu_2|.$$

Indeed, we first observe that $(t|\mu_1| + (1-t)|\mu_2|)(\Theta) = 1$ and

$$\delta_\theta = |\delta_\theta| \leq t|\mu_1| + (1-t)|\mu_2|.$$

Then, for every Borel set $E \subseteq \Theta$, if $\theta \in E$

$$1 = \delta_\theta(E) \leq (t|\mu_1| + (1-t)|\mu_2|)(E) \leq 1,$$

and if $\theta \in \Theta \setminus E$

$$(t|\mu_1| + (1-t)|\mu_2|)(E) = (t|\mu_1| + (1-t)|\mu_2|)(\Theta) - (t|\mu_1| + (1-t)|\mu_2|)(\Theta \setminus E) = 0.$$

Therefore, $|\mu_1| = |\mu_2| = \delta_\theta$, which implies $\mu_1 = \alpha_1\delta_\theta$ and $\mu_2 = \alpha_2\delta_\theta$ with $|\alpha_1| = |\alpha_2| = 1$, and equation (67) becomes

$$\alpha\delta_\theta = (t\alpha_1 + (1-t)\alpha_2)\delta_\theta. \tag{68}$$

Since $\alpha, \alpha_1, \alpha_2 \in \{-1, 1\}$, equation (68) is satisfied if and only if $\alpha = \alpha_1 = \alpha_2$. So that, $\alpha\delta_\theta = \mu_1 = \mu_2$, and then $\alpha\delta_\theta \in \operatorname{Ext}(B)$. It remains to prove the opposite inclusion $\operatorname{Ext}(B) \subseteq \{\pm\delta_\theta : \theta \in \Theta\}$. We suppose that there exists $\mu \in \mathcal{M}(\Theta)$ such that $\mu \notin \{\pm\delta_\theta : \theta \in \Theta\}$ but $\mu \in \operatorname{Ext}(B)$. Then, $\|\mu\|_{\mathrm{TV}} = 1$. We denote by χ_E the indicator function on a subset $E \subseteq \Theta$. For every Borelian set E such that $|\mu|(E) \in (0,1)$, we can rewrite μ as the linear combination

$$\mu = \mu \cdot \chi_E + \mu \cdot \chi_{\Theta\setminus E} = t\frac{\mu \cdot \chi_E}{|\mu|(E)} + (1-t)\frac{\mu \cdot \chi_{\Theta\setminus E}}{|\mu|(\Theta \setminus E)},$$

where $t = |\mu|(E) \in (0,1)$. Since $\mu \notin \{\pm\delta_\theta : \theta \in \Theta\}$, then it is possible to find a Borelian set E such that $\mu \neq |\mu|(E)^{-1}\mu \cdot \chi_E$ and $\mu \neq |\mu|(\Theta \setminus E)^{-1}\mu \cdot \chi_{\Theta\setminus E}$. This shows that there exist $t \in (0,1)$, $\mu_1, \mu_2 \in B$ such that $\mu = t\mu_1 + (1-t)\mu_2$, which yields a contradiction. Therefore, we have shown that $\operatorname{Ext}(B_{\mathrm{TV}}(1)) \subseteq \{\pm\delta_\theta : \theta \in \Theta\}$, which concludes the proof. □

To establish our representer theorem we recall the following known result.

Theorem B.3 *([8, Theorem 3.3]). Consider the problem*

$$\inf_{u\in U} F(\mathcal{A}u) + G(u), \tag{69}$$

where U is a locally convex topological vector space, $\mathcal{A} : U \to H$ is a continuous, surjective linear map with values in a finite-dimensional Hilbert space H, $F : H \to (-\infty, +\infty]$ is proper, convex, coercive and lower semi-continuous, and $G : U \to [0, +\infty)$ is a coercive and lower semi-continuous norm. Then (69) has solutions of the form $\sum_{i=1}^K \gamma_i u_i$ with $K \leq \dim H$, $\gamma_i > 0$, $\sum_{i=1}^K \gamma_i = G(u)$, and $u_i \in \operatorname{Ext}(\{u \in U : G(u) \leq 1\})$.

Theorem B.3 is a simplified version of [8, Theorem 3.3], where G is only assumed to be a seminorm. In such a case, the statement needs to take care of the kernel of G. A seminorm G is called coercive if, for all $R > 0$, the set

$$\{[u] \in U/\mathcal{N} : G(u) \leq R\}$$

is compact in $U/\mathcal{N}$, where $\mathcal{N}$ is the kernel of G (see Assumption [H1] in [8]).

Remark B.4. In Theorem B.3, the space U is endowed with a topology weaker than the topology induced by the norm G in order to ensure that the closed balls are compact.

References

[1] N. Aronszajn, Theory of reproducing kernels, Trans. Am. Math. Soc. 68 (1950) 337–404.

[2] S. Aziznejad, M. Unser, Multikernel regression with sparsity constraint, SIAM J. Math. Data Sci. 3 (1) (2021) 201–224.

[3] F. Bach, Breaking the curse of dimensionality with convex neural networks, J. Mach. Learn. Res. 18 (19) (2017) 1–53.

[4] A.R. Barron, Approximation and estimation bounds for artificial neural networks, Mach. Learn. 14 (1) (1994) 115–133.

[5] A.R. Barron, Universal approximation bounds for superpositions of a sigmoidal function, IEEE Trans. Inf. Theory 39 (3) (1993) 930–945.

[6] M. Belkin, D. Hsu, S. Ma, S. Mandal, Reconciling modern machine-learning practice and the classical bias-variance trade-off, Proc. Natl. Acad. Sci. 116 (32) (2019) 15849–15854.

[7] C. Boyer, A. Chambolle, Y.D. Castro, V. Duval, F. De Gournay, P. Weiss, On representer theorems and convex regularization, SIAM J. Optim. 29 (2) (2019) 1260–1281.

[8] K. Bredies, M. Carioni, Sparsity of solutions for variational inverse problems with finite-dimensional data, Calc. Var. Partial Differ. Equ. 59 (14) (2020).

[9] H. Brezis, Functional Analysis, Sobolev Spaces and Partial Differential Equations, Universitext, Springer, New York, 2011, pp. xiv+599.

[10] C. Carmeli, E. De Vito, A. Toigo, V. Umanitá, Vector valued reproducing kernel Hilbert spaces and universality, Anal. Appl. 8 (01) (2010) 19–61.

[11] S.S. Chen, D.L. Donoho, M.A. Saunders, Atomic decomposition by basis pursuit, SIAM Rev. 43 (1) (2001) 129–159.

[12] L. Chizat, F. Bach, Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss, in: Conference on Learning Theory, PMLR, 2020, pp. 1305–1338.

[13] P.L. Combettes, S. Salzo, S. Villa, Regularized learning schemes in feature Banach spaces, Anal. Appl. 16 (1) (2018) 1–54.

[14] F. Cucker, S. Smale, On the mathematical foundations of learning, Bull. Am. Math. Soc. 39 (2002) 1–49.

[15] G. Cybenko, Approximation by superpositions of a sigmoidal function, Math. Control Signals Syst. 2 (1989) 303–314.

[16] S. Fisher, J.W. Jerome, Spline solutions to L1 extremal problems in one and several variables, J. Approx. Theory 13 (1) (1975) 73–83.

[17] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.

[18] R. Gribonval, G. Kutyniok, M. Nielsen, F. Voigtlaender, Approximation spaces of deep neural networks, in: Constructive Approximation, 2021, pp. 1–109.

[19] L. Györfi, M. Kohler, A. Krzyzak, H. Walk, A Distribution-Free Theory of Nonparametric Regression, Springer, 2002.

[20] S. Helgason, The Radon transform on Euclidean spaces, compact two-point homogeneous spaces and Grassmann manifolds, Acta Math. 113 (1965) 153–180.

[21] S. Helgason, The Radon Transform, second edn., Progress in Mathematics, vol. 5, Birkhäuser Boston, Inc., Boston, MA, 1999.

[22] A. Hertle, On the range of the Radon transform and its dual, Math. Ann. 267 (1) (1984) 91–99.

[23] M. Holschneider, Wavelets. An Analysis Tool, Oxford Mathematical Monographs, The Clarendon Press, Oxford University Press, New York, 1995.

[24] A.A. Kilbas, O. Marichev, S. Samko, Fractional integrals and derivatives, Theory Appl. (1993).

[25] G.S. Kimeldorf, G. Wahba, A correspondence between Bayesian estimation on stochastic processes and smoothing by splines, Ann. Math. Stat. 41 (2) (1970) 495–502.

[26] G. Kimeldorf, G. Wahba, Some results on Tchebycheffian spline functions, J. Math. Anal. Appl. 33 (1) (1971) 82–95.

[27] S. Kostadinova, S. Pilipović, K. Saneva, J. Vindas, The ridgelet transform of distributions, Integral Transforms Spec. Funct. 25 (5) (2014) 344–358.

[28] A. Krogh, J. Hertz, A simple weight decay can improve generalization, in: J. Moody, S. Hanson, R.P. Lippmann (Eds.), Advances in Neural Information Processing Systems, vol. 4, Morgan-Kaufmann, 1992.

[29] R. Lin, H. Zhang, J. Zhang, On reproducing kernel Banach spaces: generic definitions and unified framework of constructions, arXiv:1901.01002, 2019.

[30] D. Ludwig, The Radon transform on Euclidean space, Commun. Pure Appl. Math. 19 (1) (1966) 49–81.

[31] C.A. Micchelli, Y. Xu, H. Zhang, Universal kernels, J. Mach. Learn. Res. 7 (95) (2006) 2651–2667.

[32] K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf, Kernel mean embedding of distributions: a review and beyond, preprint, arXiv:1605.09522, 2016.

[33] F. Natterer, The Mathematics of Computerized Tomography, SIAM, 2001.

[34] R.M. Neal, Bayesian Learning for Neural Networks, vol. 118, Springer Science & Business Media, 2012.

[35] B. Neyshabur, R.R. Salakhutdinov, N. Srebro, Path-SGD: path-normalized optimization in deep neural networks, in: C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 28, Curran Associates, Inc., 2015.

[36] G. Ongie, R. Willett, D. Soudry, N. Srebro, A function space view of bounded norm infinite width ReLU nets: the multivariate case, arXiv:1910.01635, 2019.

[37] R. Parhi, R.D. Nowak, Banach space representer theorems for neural networks and ridge splines, J. Mach. Learn. Res. 22 (43) (2021) 1–40.

[38] A. Pinkus, Approximation theory of the MLP model in neural networks, Acta Numer. 8 (1999) 143–195.

[39] Q. Que, M. Belkin, Back to the future: radial basis function networks revisited, in: Artificial Intelligence and Statistics, PMLR, 2016, pp. 1375–1383.

[40] A. Rahimi, B. Recht, Random features for large-scale kernel machines, Adv. Neural Inf. Process. Syst. 20 (2008).

[41] S. Rosset, G. Swirszcz, N. Srebro, J. Zhu, ℓ_1 Regularization in infinite dimensional feature spaces, in: International Conference on Computational Learning Theory, Springer, 2007, pp. 544–558.

[42] S. Rosset, J. Zhu, T. Hastie, Boosting as a regularized path to a maximum margin classifier, J. Mach. Learn. Res. 5 (2004) 941–973.

[43] A. Rudi, L. Rosasco, Generalization properties of learning with random features, in: Conference on Neural Information Processing Systems, vol. 31, 2017, pp. 3215–3225.

[44] W. Rudin, Functional Analysis, second edn., International Series in Pure and Applied Mathematics, McGraw-Hill, Inc., New York, 1991, pp. xviii+424.

[45] P. Savarese, I. Evron, D. Soudry, N. Srebro, How do infinite width bounded norm networks look in function space?, in: Conference on Learning Theory, PMLR, 2019, pp. 2667–2690.

[46] B. Schölkopf, R. Herbrich, A.J. Smola, A generalized representer theorem, in: International Conference on Computational Learning Theory, Springer, 2001, pp. 416–426.

[47] B. Schölkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, 2002.

[48] L. Schwartz, Théorie des distributions, in: Publications de l'Institut de Mathématique de l'Université de Strasbourg, No. IX-X. Nouvelle édition, entièrement corrigée, refondue et augmentée, Hermann, Paris, 1966, pp. xiii+420.

[49] G. Song, H. Zhang, Reproducing kernel Banach spaces with the ℓ^1 norm II: error analysis for regularized least square regression, Neural Comput. 23 (10) (2011) 2713–2729.

[50] G. Song, H. Zhang, F.J. Hickernell, Reproducing kernel Banach spaces with the ℓ^1 norm, Appl. Comput. Harmon. Anal. 34 (1) (2013) 96–116.

[51] I. Steinwart, A. Christmann, Support Vector Machines, Springer-Verlag, New York, 2008.

[52] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. R. Stat. Soc. B 58 (1) (1996) 267–288.

[53] H. Triebel, Theory of Function Spaces, Modern Birkhäuser Classics, Birkhäuser/Springer, Basel AG, Basel, 2010, p. 285.

[54] M. Unser, A unifying representer theorem for inverse problems and machine learning, Found. Comput. Math. (2020) 1–20.

[55] M. Unser, J. Fageot, J.P. Ward, Splines are universal solutions of linear inverse problems with generalized TV regularization, SIAM Rev. 59 (4) (2017) 769–793.

[56] V. Vapnik, Statistical Learning Theory, Wiley, 1998.

[57] H. Wendland, Scattered Data Approximation, vol. 17, Cambridge University Press, 2004.

[58] Y. Xu, Q. Ye, Generalized Mercer Kernels and Reproducing Kernel Banach Spaces, vol. 258, American Mathematical Society, 2019.

[59] C. Zhang, S. Bengio, M. Hard, B. Recht, O. Vinyals, Understanding deep learning requires rethinking generalization, in: International Conference on Learning Representations, 2017, pp. 1–15.

[60] H. Zhang, Y. Xu, J. Zhang, Reproducing kernel Banach spaces for machine learning, J. Mach. Learn. Res. 10 (95) (2009) 2741–2775.

[61] S. Zuhovickii, Remarks on problems in approximation theory, Mat. Zbirnik KDU (1948) 169–183.