# NEURAL REPRODUCING KERNEL BANACH SPACES AND REPRESENTER THEOREMS FOR DEEP NETWORKS

F. BARTOLUCCI, E. DE VITO, L. ROSASCO, AND S. VIGOGNA

ABSTRACT. Characterizing the function spaces defined by neural networks helps understanding the corresponding learning models and their inductive bias. While in some limits neural networks correspond to function spaces that are Hilbert spaces, these regimes do not capture the properties of the networks used in practice. Indeed, several results have shown that shallow networks can be better characterized in terms of suitable Banach spaces. However, analogous results for deep networks are limited. In this paper we show that deep neural networks define suitable reproducing kernel Banach spaces. These spaces are equipped with norms that enforce a form of sparsity, enabling them to adapt to potential latent structures within the input data and their representations. In particular, by leveraging the theory of reproducing kernel Banach spaces, combined with variational results, we derive representer theorems that justify the finite architectures commonly employed in applications. Our study extends analogous results for shallow networks and represents a step towards understanding the function spaces induced by neural architectures used in practice.

## 1. INTRODUCTION

Neural networks define functions by composing linear and nonlinear maps in a multi-layer (deep) architecture. While easy to implement, the corresponding models are hard to analyze since they are nonlinearly parameterized. The study of the function spaces defined by different neural network architectures provides insights into the corresponding learning models. In particular, it provides indications on the underlying inductive bias, namely, which functions can be approximated and learned efficiently by a given class of networks.

In some over-parameterized regimes, neural networks can be seen to define Hilbert spaces of functions and in particular reproducing kernel Hilbert spaces (RKHS) [1]. It is a classical observation that shallow networks with infinitely many random units correspond to RKHS, with reproducing kernels depending on the considered activation function [19]. This regime, also known as the Gaussian Process (GP) limit, has connections with models such as random features [23]. The limits of more complex, possibly non-shallow architectures can also be derived and characterized in terms of RKHS, see e.g. [16]. Another infinite-width limit in which neural networks are described by RKHS is the so-called lazy training regime [10]. In this limit, the network weights evolve little during the optimization and can be well approximated by a linear approximation around a random initialization. Also in this case, the corresponding function spaces are RKHS, and the associated kernel is called neural tangent kernel (NTK) [17]. Again, neural tangent kernels and corresponding RKHS can be derived for a variety of architectures, see e.g. [5].

However, neither the above GP/kernel limit nor the NTK/lazy training regime appear to capture key aspects of neural network models [15, 4] used in practice. Results for shallow networks suggest that neural networks might favor functions with small norms that are not Hilbertian but rather associated with Banach spaces [2]. In turn, representer theorems associated to such norms allow to derive finite-width networks commonly used in practice from variational principles [26, 20, 28, 21, 22]. These observations have sparked interest in understanding Banach spaces associated to neural networks. One possibility is to consider extensions of classical splines [32, 31, 30, 29]. Another possibility is to consider reproducing kernel Banach spaces (RKBS) [34, 18], see e.g. [3] and references therein.

In this paper, we develop the latter perspective tackling the extension from shallow to deep networks. The study of Banach spaces associated to deep architectures and corresponding representer theorems was started in [22], where deep architectures with ReLU activations and finite rank constraints at each layer are considered. The latter requirement is not natural and is mainly due to technical reasons. Indeed, finite rank constraints allow for the construction of layers as concatenation of vector valued functions studied for shallow networks. In our study, we propose an approach that avoids finite rank constraints and allows to consider more general activations. This requires more substantial developments employing vector measures of finite variation to address the challenges posed by potentially infinite-dimensional hidden layers. Our first contribution is to define a reproducing kernel Banach space which describes an infinite-width limit of a deep neural network with an associated norm promoting sparsity. We call such a space a neural RKBS. Then, we provide a representer theorem for a large class of nonlinearities that shows how neural networks minimizing empirical objectives can be taken to have a finite width at every layer. This result extends analogous results for shallow networks. It implies that commonly used networks are optimal in the sense that they are solutions of a suitable variational problem.

The rest of the paper is organized as follows. In Section 2 we provide background for the study of the paper. In Section 3 we review the main technical ingredients of our construction, namely reproducing kernel Banach spaces and vector Radon measures. In Section 4 we introduce deep integral RKBS to model functional properties of deep neural networks, leading to the construction of neural RKBS. In Section 5 we prove our represent theorems for deep neural networks. In Section 6, we construct a particular instance of neural RKBS with a countably infinite number of neurons per hidden layer, which permits a more explicit form of the relative representer theorem. In Appendix A we collect variational results and extreme point characterizations used to prove our representer theorems. Table 1 summarizes the main notation we use in the paper.

TABLE 1. Notation

| symbol | definition | symbol | definition |
|---|---|---|---|
| $\mathcal{X}, \mathcal{Y}$ | Banach spaces | $\Theta$ | locally compact second contable space |
| $B(\mathcal{X}, \mathcal{Y})$ | bounded linear maps $\mathcal{X} \to \mathcal{Y}$ | $\mathcal{B}(\Theta)$ | Borel $\sigma$-algebra on $\Theta$ |
| $\mathcal{X}'$ | continuous dual of $\mathcal{X}$ | $\mathcal{C}_0(\Theta, \mathcal{Y})$ | continuous functions $\Theta \to \mathcal{Y}$ vanishing at $\infty$ |
| ${}_{\mathcal{X}}\langle \cdot, \cdot \rangle_{\mathcal{X}'}$ | pairing on $\mathcal{X}, \mathcal{X}'$ | $\mathcal{C}_0(\Theta)$ | $\mathcal{C}_0(\Theta, \mathbb{R})$ |
| $\langle \cdot, \cdot \rangle_{\mathcal{X}}$ | inner product on the Hilbert space $\mathcal{X}$ | $\mathcal{M}(\Theta, \mathcal{Y})$ | vector measures on $\Theta$ with values in $\mathcal{Y}$ |
| $\| \cdot \|_{\mathcal{X}}$ | norm on $\mathcal{X}$ | $\mathcal{M}(\Theta)$ | $\mathcal{M}(\Theta, \mathbb{R})$ |
| $B_{\mathcal{X}}(r)$ | ball on $\mathcal{X}$ of radius $r$ | $\delta_\theta$ | Dirac delta at $\theta$ |
| $\mathrm{Ext}(Q)$ | extremal points of $Q \subset \mathcal{X}$ | $\| \cdot \|_{\mathrm{TV}}$ | total variation norm |

## 2. Preliminary Discussion

In this section, we provide the necessary background by introducing the notation for deep neural networks, reviewing the infinite-width limit of shallow networks, and presenting a high-level overview of the proposed infinite-width limit for deep architectures.

**Neural networks.** We start by setting up some notation and introduce fully connected feed-forward neural networks.

**Definition 2.1** (Fully connected feed-forward neural network). *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a (nonlinear) function, $L \geq 1$ an integer and*

$$d = d_0, d_1, \dots, d_L, d_{L+1} = p \geq 1$$

*a family of $L+2$ integers. A function $f : \mathbb{R}^d \to \mathbb{R}^p$ is called a fully connected feed-forward neural network from $\mathbb{R}^d$ to $\mathbb{R}^p$ with activation function $\sigma$, depth $L$ and widths $d_1, \dots, d_L$ if*

$$f(x) = x^{(L+1)}$$

*where, for each $x \in \mathbb{R}^d$, the vector $x^{(L+1)} \in \mathbb{R}^p$ is defined by the following recursive equation*

$$\begin{cases} x^{(1)} = W^{(1)} x + b^{(1)} & \in \mathbb{R}^{d_1} \\ x^{(\ell+1)} = W^{(\ell+1)} \sigma(x^{(\ell)}) + b^{(\ell+1)} & \in \mathbb{R}^{d_{\ell+1}}, \qquad \ell = 1, \dots, L, \end{cases} \tag{1}$$

*for some weights $W^{(\ell)} \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ and biases (or offsets) $b^{(\ell)} \in \mathbb{R}^{d_\ell}$. We call a neural network shallow (or a one-hidden layer network) if $L = 1$, deep if $L > 1$.*

In (1) the activation function $\sigma$ is assumed to be applied on vectors component-wise, and the vector $\sigma(x^{(\ell)})$ for $\ell = 1, \dots, L$ is the $\ell$-th *hidden layer* of the network. The input and output dimensions $d = d_0, p = d_{L+1}$ are fixed by the problem, and $d_\ell$ is the number of neurons (or units) at the $\ell$-th hidden layer. A *neuron* is a function of the form $\phi(z) = \sigma(\langle z, w \rangle + b)$. Hence, the definition of a network requires specifying an activation function $\sigma$, a depth $L$, and the widths $d_1, \dots, d_L$. These parameters define the network architecture. Given an architecture, the parameters $W^{(\ell)}, b^{(\ell)}$ in $\mathbb{R}^{d_\ell \times d_{\ell-1}} \times \mathbb{R}^{d_\ell}$ are found optimizing an empirical objective function.

The neural networks with a fixed architecture form a subset $\mathcal{F}_{NN}$ of functions from $\mathbb{R}^d$ to $\mathbb{R}^p$ parameterized nonlinearly by weights and biases. This is already clear considering scalar-valued shallow neural networks of width $K \in \mathbb{N}$, that is

$$\mathcal{F}_{NN} = \left\{ f : \mathbb{R}^d \to \mathbb{R} : f(x) = \sum_{k=1}^{K} v_k \sigma(w_k^\top x + b_k) + b, w_k \in \mathbb{R}^d, v_k, b, b_k \in \mathbb{R} \right\}. \tag{2}$$

The nonlinear dependence on the parameters is a key challenge in defining and characterizing the corresponding function spaces. For shallow networks, considering the so called infinite-width limit provides an approach to tackle this question.

**Infinite-width limit of shallow neural networks.** In the context of shallow networks, the infinite-width limit corresponds to consider functions parameterized by measures rather than weights, that is

$$f_\mu(x) = \int_{\mathbb{R}^d \times \mathbb{R}} \sigma(w^\top x + b) \, \mathrm{d}\mu(w, b), \qquad \mu \in \mathcal{M}(\mathbb{R}^d \times \mathbb{R}). \tag{3}$$

Generalizing further, one can consider an arbitrary input space $\mathcal{X}$ and functions $f$ from $\mathcal{X}$ to $\mathbb{R}$ of the form

$$f_\mu(x) = \int_\Theta \rho(x,\theta)\, \mathrm{d}\mu(\theta), \qquad \mu \in \mathcal{M}(\Theta). \tag{4}$$

for a suitable locally compact second countable parameter space $\Theta$ and basis function $\rho : \mathcal{X} \times \Theta \to \mathbb{R}$.

We add two observations. First, the expression in (3) needs some care. Indeed, commonly used activation functions are not integrable, and typical parameter spaces $\Theta$ are non-compact, so that the integrals in (3) (and (4)) may not converge. For example, in [3] the activation is multiplied by a smoothing function to ensure integrability. The corresponding function space is shown to be a reproducing kernel Banach space, with norm induced by the total variation norm on measures. Second, the finite shallow networks in the set (2) can be recovered from the infinite-width model (3) by taking finite linear combinations of Dirac deltas as measures. Interestingly, finite networks can be shown to emerge from variational principles. Specifically, if the minimization of an empirical objective is performed over measures, instead of weights, then the optimal solutions are atomic measures. In other words, finite networks are optimal solutions of empirical minimization problems over possibly infinite dimensional networks. We refer for example to [29] for an account of recent works addressing the above topics. As discussed next, our main motivation is studying similar questions in the context of deep networks.

**Infinite-width limit of deep neural networks.** Previous work on Banach spaces associated with deep architectures, and corresponding representation theorems, focus on layers which are concatenation of of infinite-width shallow neural networks with finite-dimensional outputs. As a result, this approach inherently restricts the deep architecture to finite-rank constraints at every layer [22]. In contrast, our work presents an alternative framework that avoids such constraints and accommodates a broader class of activation functions. This generalization requires addressing the challenges posed by potentially infinite-dimensional hidden layers. To this end, we construct deep architectures as compositions of RKBS of integrable functions valued in infinite-dimensional spaces. More precisely, we begin by considering the direct sum of integral RKBS, forming a linear space, and use this structure to parameterize spaces of composed functions, thereby yielding a corresponding nonlinear function space, which we call a deep integral RKBS. By further specializing this construction to integral functions defined through activation functions, we arrive at what we call a neural RKBS. This formulation provides a natural and rigorous connection with commonly used deep neural network models. Finally, we use these function spaces to derive novel representer theorems for deep networks. These representer theorems characterize the minimization of empirical objective functions over deep RKBS. In particular, they show that finite networks are optimal from a variational perspective. Next, we develop these ideas in detail.

## 3. Vector measures, integral RKBS and Neural Networks

We introduce the following definition, that readily generalizes vector-valued reproducing kernel Hilbert spaces [8] to a Banach setting. We refer to [34, 18] for an overview.

**Definition 3.1** (Vector-valued RKBS). *Let $\mathcal{X}$ be a set and $\mathcal{Y}$ a Banach space. A* reproducing kernel Banach space *(RKBS) $\mathcal{H}$ on $\mathcal{X}$ with values in $\mathcal{Y}$ is a Banach space such that*

(i) *the elements of $\mathcal{H}$ are functions $f : \mathcal{X} \to \mathcal{Y}$;*

(ii) *the sum and the multiplication by scalars in $\mathcal{H}$ are defined pointwise;*

(iii) *for all $x \in \mathcal{X}$ there is $C_x > 0$ such that $\|f(x)\|_{\mathcal{Y}} \leq C_x \|f\|_{\mathcal{H}}$ for all $f \in \mathcal{H}$.*

The first two conditions are equivalent to say that $\mathcal{H}$ is a subspace of $\mathcal{Y}^{\mathcal{X}}$, the vector space of all functions from $\mathcal{X}$ to $\mathcal{Y}$, while the third condition requires that for all $x \in \mathcal{X}$ the pointwise evaluation $f \mapsto f(x)$ is in $B(\mathcal{H}, \mathcal{Y})$, the space of bounded linear operators between $\mathcal{H}$ and $\mathcal{Y}$.

Reproducing kernel Hilbert spaces can be characterized in terms of so-called feature maps. In the following proposition we provide an analogous result for RKBS. In this case, feature spaces are Banach spaces.

**Proposition 3.2.** *Let $\mathcal{X}$ be a set, $\mathcal{Y}$ a Banach space and $\mathcal{H}$ a set of functions $f : \mathcal{X} \to \mathcal{Y}$. Consider the following statements.*

(a) *The space $\mathcal{H}$ is a RKBS.*

(b) *There is a Banach space $\mathcal{F}$ and a map $\phi : \mathcal{X} \to B(\mathcal{F}, \mathcal{Y})$ such that*

  (i) $\mathcal{H} = \{f_\mu : \mu \in \mathcal{F}\}$ *where* $f_\mu = \phi(\cdot)\mu$ ;

  (ii) $\|f\|_{\mathcal{H}} = \inf\{\|\mu\|_{\mathcal{F}} : \mu \in \mathcal{F}, f = f_\mu\}$ .

(c) *There is a Banach space $\mathcal{F}$ and a map $\psi : \mathcal{X} \to B(\mathcal{Y}', \mathcal{F}')$ such that*

  (i) $\mathcal{H} = \{f_\mu : \mu \in \mathcal{F}\}$ *where* ${}_{\mathcal{Y}''}\langle f_\mu(\cdot), y' \rangle_{\mathcal{Y}'} = {}_{\mathcal{F}}\langle \mu, \psi(\cdot)y' \rangle_{\mathcal{F}'}$ *for all* $y' \in \mathcal{Y}'$;

  (ii) $\|f\|_{\mathcal{H}} = \inf\{\|\mu\|_{\mathcal{F}} : \mu \in \mathcal{F}, f = f_\mu\}$ .

*Then* (a) *and* (b) *are equivalent and each one implies* (c). *Moreover, if $\mathcal{Y}$ is reflexive (in particular Hilbert), then* (a), (b) *and* (c) *are all equivalent.*

*Proof.* To see that (a) implies (b), take $\mathcal{F} = \mathcal{H}$ and define

$$\phi : \mathcal{X} \to B(\mathcal{F}, \mathcal{Y}), \qquad \phi(x)f = f(x).$$

Then (i) and (ii) of item (b) are clear. Let us prove that (b) implies (a). Clearly $\mathcal{H}$ is a linear space and $\|\cdot\|_{\mathcal{H}}$ is a norm. We then show that the normed space $\mathcal{H}$ is complete. The linear map $\mu \mapsto f_\mu$ has kernel $\mathcal{N} = \bigcap_{x \in \mathcal{X}} \ker \phi(x)$. Since $\phi(x)$ is bounded for all $x \in \mathcal{X}$, $\ker \phi(x)$ is closed, hence so is $\mathcal{N}$. Thus, $\mathcal{F}/\mathcal{N}$ is a Banach space [24, Theorem 1.41] isomorphic to $\mathcal{H}$ by construction, which is therefore complete. Next, we show that point evaluations in $\mathcal{H}$ are continuous. For $f \in \mathcal{H}$, let $\mu \in \mathcal{F}$ such that $f = f_\mu$. Then

$$\|f(x)\|_{\mathcal{Y}} = \|\phi(x)\mu\|_{\mathcal{Y}} \leq \|\phi(x)\|_{B(\mathcal{F},\mathcal{Y})} \|\mu\|_{\mathcal{F}},$$

whence

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

The implication from (b) to (c) follows easily considering $\psi(x) = \phi(x)^t \in \mathcal{B}(\mathcal{Y}', \mathcal{F}')$ be the transpose map of $\phi(x)$. Finally, note that (i) in (c) defines $f_\mu$ as a function from $\mathcal{X}$ to $\mathcal{Y}''$. Hence, if $\mathcal{Y}$ is reflexive, it defines $f_\mu : \mathcal{X} \to \mathcal{Y}$. From here, following the proof of the implication from (b) to (a), one can prove that (c) implies (a). □

**Vector Radon measures.** In view of Proposition 3.2, a RKBS can be constructed choosing a suitable feature space. Thinking of a neural network layer as an atomic integration, we will define RKBS parameterized by measures. Since layers have vectorial outputs, we need the notion of vector valued measure [11].

Let $\Theta$ be a Hausdorff, locally compact, second countable topological space, and let $\mathcal{Y}$ be a Banach space. Recall that a (numerable) partition of a Borel set $A$ is a numerable family of Borel sets $\{A_i\}$ such that $A_i \cap A_j = \emptyset$ for all $i \neq j$ and $\bigcup_i A_i = A$.

**Definition 3.3** (Vector measure). *A* vector measure *on $\Theta$ with values in $\mathcal{Y}$ is a set function $\mu \colon \mathcal{B}(\Theta) \to \mathcal{Y}$ such that for all $A \in \mathcal{B}(\Theta)$ and all $\{A_i\}$ partitions of $A$,*

$$\mu(A) = \sum_i \mu(A_i),$$

*where the sum converges unconditionally in the $\| \cdot \|_{\mathcal{Y}}$-norm.*

Pettis theorem [14, Thm.1 IV.10.1] shows that a set function $\mu \colon \mathcal{B}(\Theta) \to \mathcal{Y}$ is a vector measure if and only if $\mu_{y'} = {}_{\mathcal{Y}}\langle \mu(\cdot), y' \rangle_{\mathcal{Y}'}$ is a $\sigma$-additive measure for all $y' \in \mathcal{Y}'$.

**Definition 3.4** (Variation of a vector measure). *Let $\mu$ be a vector measure on $\Theta$ with values in $\mathcal{Y}$. The variation of $\mu$ is the function $|\mu| \colon \mathcal{B}(\Theta) \to [0, +\infty]$ defined by*

$$|\mu|(A) = \sup_{\{A_i\}} \sum_i \|\mu(A_i)\|_{\mathcal{Y}} \qquad A \in \mathcal{B}(\Theta),$$

*where the supremum is taken over all finite partitions of $A$. If $|\mu|(\Theta) < +\infty$, the measure $\mu$ is called a vector measure of bounded variation.*

The space $\mathcal{M}(\Theta, \mathcal{Y})$ of vector measures of bounded variation is a Banach space with respect to the norm

$$\|\mu\|_{\mathrm{TV}} = |\mu|(\Theta).$$

If $\mu \in \mathcal{M}(\Theta, \mathcal{Y})$, its variation $|\mu|$ is a finite positive measure on $\Theta$, see [13, 1.A.10].

The integration of a scalar function $\varphi$ with respect a vector measure of bounded variation can be defined as the Bochner integral of a vector valued function, see [13, Ch.1 Section D] and [14, p. IV.10]. In particular, a measurable scalar function $\varphi$ is integrable with respect to $\mu$ if and only if $\varphi$ is integrable with respect to $|\mu|$. If $\varphi = \sum_i t_i \chi_{A_i}$ is a simple function, then

$$\int_\Theta \varphi(\theta) d\mu(\theta) = \sum_i t_i \mu(A_i) \in \mathcal{Y},$$

and the integral of an arbitrary $|\mu|$-integrable functions is defined via the density of simple functions.

If $\mathcal{Y}$ has the Radon-Nikodym property [13, Ch. 1.G], as it happens if $\mathcal{Y}$ is reflexive, then $\mu$ has density with respect to $|\mu|$, *i.e.* there exists a function $g \colon \Theta \to \mathcal{Y}$ such that $\|g(\theta)\|_{\mathcal{Y}} = 1$ for all $\theta \in \Theta$, $g$ is Bocnher integrable with respect to $|\mu|$, and

$$\int_\Theta \varphi(\theta) d\mu(\theta) = \int_\Theta \varphi(\theta) g(\theta) d|\mu|(\theta),$$

for all $\mu$-integrable scalar functions $\varphi : \Theta \to \mathbb{R}$. If $\mathcal{Y}$ does not have the Radon-Nikodym property, nevertheless there always exists a function $g : \Theta \to \mathcal{Y}$ such that

i) for all $y' \in \mathcal{Y}'$, ${}_{\mathcal{Y}}\langle g(\cdot), y' \rangle_{\mathcal{Y}'}$ is $|\mu|$-integrable;
ii) for all $\theta \in \Theta$, $\|g(\theta)\|_{\mathcal{Y}} = 1$;
iii) for all scalar function $\varphi$ that are $\mu$-integrable and $y' \in \mathcal{Y}'$

$${}_{\mathcal{Y}}\langle \int_\Theta \varphi(\theta) d\mu(\theta), y' \rangle_{\mathcal{Y}'} = \int \varphi(\theta) {}_{\mathcal{Y}}\langle g(\theta), y' \rangle_{\mathcal{Y}'} d|\mu|(\theta),$$

see [13, Ch.1 Theorem 34].

**Vector integral RKBS and neural networks.** We now introduce particular classes of RKBS. In particular, we extend the infinite-width limit (4) of shallow neural networks from scalar to vector-valued functions. This result is of independent interest and will be crucial for the extension from shallow to deep networks.

**Definition 3.5** (Integral RKBS). *Let $\Theta$ be a locally compact, second countable topological space, $\mathcal{X}$ a set, and $\mathcal{Y}$ a Banach space. Let*

$$\rho : \mathcal{X} \times \Theta \to \mathbb{R}$$

*be such that $\rho(x, \cdot) \in C_0(\Theta)$ for all $x \in \mathcal{X}$. For $\mu \in \mathcal{M}(\Theta, \mathcal{Y})$, let*

$$f_\mu : \mathcal{X} \to \mathcal{Y}, \qquad f_\mu(x) = \phi(x)\mu = \int_\Theta \rho(x, \theta) \mathrm{d}\mu(\theta), \tag{5}$$

*and*

$$\mathcal{H} = \{f_\mu : \mathcal{X} \to \mathcal{Y} : \mu \in \mathcal{M}(\Theta, \mathcal{Y})\},$$

*with*

$$\|f\|_{\mathcal{H}} = \inf_{\mu \in \mathcal{M}(\Theta, \mathcal{Y})} \{\|\mu\|_{\mathrm{TV}} : f = f_\mu\}. \tag{6}$$

*The space $\mathcal{H}$ thus defined is a (vector-valued) RKBS, which we call an* integral RKBS. *Moreover, we call $\Theta$ the* parameter spaces *of $\mathcal{H}$, and $\rho$ its* basis function.

**Remark 3.6.** *If $\rho(x, \cdot)$ is bounded for all $x \in \mathcal{X}$,* (5) *is still well defined. The stronger assumption that $\rho(x, \cdot) \in C_0(\Theta)$ ensures that the map $f \mapsto f(x)$ is weakly* continuous, see Remark 5.2.*

**Remark 3.7.** *Scalar valued integral RKBS correspond to the choice $\mathcal{Y} = \mathbb{R}$ and coincide with the setting considered in [3].*

## 4. Deep Integral and Neural RKBS

In this section, we start introducing *deep* integral RKBS. Then, we derive representer theorems on these spaces, which show how optimal solutions minimizing empirical objectives can be taken to have a finite width at every layer.

**Deep RKBS.** The first step is to take direct sums of RKBS and define the deep RKBS as a composition of elements in the direct sum.

**Definition 4.1** (Deep RKBS). *Let $\mathcal{X}$ be a set and $\mathcal{Y}$ a Banach space. Fix a positive integer $L \geq 1$. Take a set $\mathcal{X}_0 = \mathcal{X}$ and Banach spaces $\mathcal{X}_1, \dots, \mathcal{X}_{L+1} = \mathcal{Y}$. For $\ell = 0, \dots, L$, take RKBS $\mathcal{H}_\ell$ on $\mathcal{X}_\ell$ with values in $\mathcal{X}_{\ell+1}$. The direct sum*

$$\mathcal{H} = \mathcal{H}_0 \oplus \cdots \oplus \mathcal{H}_L$$

*is a Banach space with respect to the norm*

$$\|f\|_{\mathcal{H}} = \|f_0\|_{\mathcal{H}_0} + \cdots + \|f_L\|_{\mathcal{H}_L}, \qquad f = f_0 \oplus \cdots \oplus f_L.$$

*To every $f = f_0 \oplus \cdots \oplus f_L \in \mathcal{H}$ we assign the function $f^{\mathrm{deep}} : \mathcal{X} \to \mathcal{Y}$ defined by*

$$f^{\mathrm{deep}} = f_L \circ \cdots \circ f_0,$$

*and we set*

$$\mathcal{H}^{\mathrm{deep}} = \{f^{\mathrm{deep}} : \mathcal{X} \to \mathcal{Y} : f \in \mathcal{H}\}$$

*endowed with the complexity measure $\Phi : \mathcal{H}^{\mathrm{deep}} \to [0, +\infty)$ given by*

$$\Phi(f^{\mathrm{deep}}) = \inf\{\|g\|_{\mathcal{H}} : g \in \mathcal{H} \text{ such that } g^{\mathrm{deep}} = f^{\mathrm{deep}}\}. \tag{7}$$

*With a slight abuse of language, we call the nonlinear space $\mathcal{H}^{\text{deep}}$ an* RKBS of depth $L$ *induced by the Banach space $\mathcal{H}$. If $L > 1$, we refer to $\mathcal{H}^{\text{deep}}$ as a* deep RKBS*. Moreover, we call $\mathcal{X}_\ell$ the* layer spaces *of $\mathcal{H}$. In particular, $\mathcal{X} = \mathcal{X}_0$ is the* input space*, $\mathcal{Y} = \mathcal{X}_{L+1}$ is the* output space*, and $\mathcal{X}_\ell$, $\ell = 1, \dots, L$, are the* hidden layer spaces*.*

**Deep integral RKBS.** Next, we combine Definitions 3.5 and 4.1.

**Definition 4.2** (Deep integral RKBS). *Let $\mathcal{H}^{\text{deep}}$ be an RKBS of depth L. If the RKBS $\mathcal{H}_\ell$ are integral for all $\ell = 0, \dots, L$,*

$$\mathcal{H}_\ell = \{f_{\mu_\ell} : \mathcal{X}_\ell \to \mathcal{X}_{\ell+1} : \mu_\ell \in \mathcal{M}(\Theta_\ell, \mathcal{X}_{\ell+1})\},$$

*with basis functions*

$$\rho_\ell : \mathcal{X}_\ell \times \Theta_\ell \to \mathbb{R},$$

*where $\rho_\ell(x, \cdot) \in \mathcal{C}_0(\Theta_\ell)$ for all $x \in \mathcal{X}$. We call $\mathcal{H}^{\text{deep}}$ an* integral RKBS of depth *$L$ and, if $L > 1$, a* deep integral RKBS*.*

Thus, a function $f^{\text{deep}} : \mathcal{X} \to \mathcal{Y}$ in a deep integral RKBS $\mathcal{H}^{\text{deep}}$ of depth $L$ has the form

$$\begin{cases} x^{(0)} = x & \in \mathcal{X} \\ x^{(\ell+1)} = \displaystyle\int_{\Theta_\ell} \rho_\ell(x^{(\ell)}, \theta_\ell) \mathrm{d}\mu_\ell(\theta_\ell) & \in \mathcal{X}_{\ell+1}, \qquad \ell = 0, \dots, L, \\ f^{\text{deep}}(x) = x^{(L+1)} & \in \mathcal{Y} \end{cases} \tag{8}$$

where $\mu_\ell \in \mathcal{M}(\Theta_\ell, \mathcal{X}^{\ell+1})$ for any $\ell = 0, \dots, L$.

We now define (deep) integral RKBS modeled on neural networks. In such spaces, the basis functions are defined in terms of an activation function.

**Definition 4.3** (Neural RKBS). *Let $\mathcal{H}^{\text{deep}}$ be an integral RKBS of depth L. Suppose that, for each $\ell$, the layer $\mathcal{X}_\ell$ is a function space over the parameter space $\Theta_\ell$, that is, $\mathcal{X}_\ell \subset \mathbb{R}^{\Theta_\ell}$. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a (nonlinear) activation function and $c_\ell : \Theta_\ell \to \mathbb{R}$. Suppose that the basis functions $\rho_\ell$ are of the form*

$$\begin{aligned} &\rho_0(x_0, \theta_0) = x_0(\theta_0) \\ &\rho_\ell(x_\ell, \theta_\ell) = \sigma(x_\ell(\theta_\ell) + c_\ell(\theta_\ell))\beta_\ell(\theta_\ell), \qquad \ell = 1, \dots, L, \end{aligned}$$

*where $\beta_\ell : \Theta_\ell \to \mathbb{R}$ is such that $\rho_\ell(x_\ell, \cdot) \in \mathcal{C}_0(\Theta_\ell)$ for all $x_\ell \in \mathcal{X}_\ell$. Then we call $\mathcal{H}^{\text{deep}}$ a* neural RKBS *of depth L and, if $L > 1$, a* deep *neural RKBS.*

**Remark 4.4** (Infinite-width shallow neural networks revisited). *In Remark 3.7, we observed that function spaces considered in [3] correspond to the choice $\mathcal{Y} = \mathbb{R}$ in Definition 3.5. Here, we show that those spaces can also be recovered from Definition 4.3 by taking $L = 1$, parameter and layer spaces*

$$\begin{aligned} \Theta_0 &= \{0, \dots, d\}, & \mathcal{X}_0 &= \mathbb{R}^d, \\ \Theta_1 &= \mathbb{R}^{d+1}, & \mathcal{X}_1 &= \mathbb{R}^{d+1}, \\ & & \mathcal{X}_2 &= \mathbb{R}\,, \end{aligned}$$

*and basis functions of the form*

$$\rho_0(x, j) = \begin{cases} 1 & j = 0 \\ x_j & j = 1, \dots, d \end{cases}, \quad x \in \mathbb{R}^d,$$

$$\rho_1(x, \theta) = \sigma(\langle x, \theta \rangle_{\mathbb{R}^{d+1}})\beta(\theta), \quad \rho_1(x, \cdot) \in \mathcal{C}_0(\mathbb{R}^d \times \mathbb{R})\,.$$

*Indeed, let* $\mu_0 \in \mathcal{M}(\Theta_0, \mathcal{X}_1) = \bigoplus_{k=0}^{d} \mathbb{R}^{d+1}$. *Then* $\mu_0 = \sum_{k=o}^{d} v_k \delta_k$ *where* $v_0, \ldots, v_d \in \mathbb{R}^{d+1}$, *so that, for all* $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$,

$$x^{(1)} = v_0 + \sum_{k=1}^{d} v_k x_k.$$

*Moreover, let* $\mu_1 \in \mathcal{M}(\Theta_1, \mathcal{X}_2) = \mathcal{M}(\mathbb{R}^{d+1})$. *Then*

$$f(x) = \int_{\mathbb{R}^{d+1}} \sigma(\sum_{k=1}^{d} \langle v_k, \theta \rangle_{\mathbb{R}^{d+1}} x_k + \langle v_0, \theta \rangle_{\mathbb{R}^{d+1}}) \beta(\theta) d\mu_1(\theta).$$

*Now, assuming that* $v_0, \ldots, v_d$ *are linearly independent, let* $\Lambda : \mathbb{R}^{d+1} \to \mathbb{R}^{d+1}$ *be the linear transformation such that* $\Lambda^\top e_k = v_k$, *when* $(e_k)_{k=0,\ldots,d}$ *is the canonical base of* $\mathbb{R}^{d+1}$. *Then, setting* $\theta = (w, b) \in \mathbb{R}^d \times \mathbb{R}$, $\beta'(\theta) = \beta(\Lambda^{-1}\theta)$, *and* $\mu$ *be the pushforward measure of* $\mu_1$ *by* $\Lambda$, *we obtain*

$$f(x) = \int_{\mathbb{R} \times \mathbb{R}^d} \sigma(\langle w, x \rangle_{\mathbb{R}^d} + b) \beta'(w, b) d\mu(w, b),$$

*These result in shallow neural networks with an uncountable number of hidden neurons. In Section 6, inspired by this construction, we will introduce deep networks where each hidden layer has countably many neurons.*

## 5. Representer Theorems for Deep Neural Networks

In this section, we state and prove representer theorems for deep integral and neural RKBS, showing that common used deep neural networks can be seen as solutions of a variational problem. We start by briefly recalling the basic supervised learning setting.

Let $\mathcal{X}$ and $\mathcal{Y}$ be sets, called input and output space, respectively. Consider a class $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathcal{Y}$, called *hypothesis space*, a *loss function* $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$, and a *penalty* $\Phi : \mathcal{H} \to [0, \infty)$. Given $N$ samples

$$(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, \qquad i = 1, \ldots, N,$$

consider the *regularized empirical risk minimization problem*

$$\min_{f \in \mathcal{H}} \mathcal{R}(f) + \Phi(f). \tag{9}$$

Here,

$$\mathcal{R}(f) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f(x_i), y_i)$$

is the empirical error associated to the loss function $\mathcal{L}$ and the training points $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i = 1, \ldots, N$. Representer theorems characterize the solutions of problems such as (9).

From now on, we fix a deep integral RKBS $\mathcal{H}^{\text{deep}}$ of depth $L$ with basis functions $\rho_\ell : \mathcal{X}_\ell \times \Theta_\ell \to \mathbb{R}$, as in Definition 4.2. To prove a representer theorem for $\mathcal{H}^{\text{deep}}$ we will need the following fundamental assumption on the basis functions $\rho_\ell$.

**Assumption 1.** *The output space* $\mathcal{Y}$ *is a Hilbert space and, for each* $\ell = 1, \ldots, L$, *the hidden layer* $\mathcal{X}_\ell$ *is a separable reproducing kernel Hilbert space on the corresponding parameter space* $\Theta_\ell$ *with a continuous reproducing kernel. Furthermore, the basis functions are of the form*

$$\rho_\ell(x, \theta) = \widetilde{\rho}_\ell(x(\theta), \theta), \qquad x \in \mathcal{X}_\ell, \quad \theta \in \Theta_\ell\,, \tag{10}$$

*where* $\widetilde{\rho}_\ell : \mathbb{R} \times \Theta_\ell \to \mathbb{R}$ *is a continuos function such that* $\rho_\ell(0, \cdot) \in \mathcal{C}_0(\Theta_\ell)$, *and*

$$|\rho_\ell(x, \theta) - \rho_\ell(x', \theta)| \le C_\ell |\langle x - x', g_\ell(\theta) \rangle_{\mathcal{X}_\ell}| |\beta_\ell(\theta)|, \qquad x, x' \in \mathcal{X}_\ell, \quad \theta \in \Theta_\ell\,, \tag{11}$$

*where $C_\ell > 0$, $g_\ell \in \mathcal{C}_b(\Theta_\ell, \mathcal{X}_\ell)$ and $\beta_\ell \in \mathcal{C}_0(\Theta_\ell)$.*

The proof of the representer theorem for deep integral RKBS (Theorem 5.3) relies on applying Theorem A.2 to each hidden layer. Theorem A.2 requires a finite-dimensional setting. For the output layer, this condition is naturally met by assuming the output space is $\mathbb{R}^p$. However, since the hidden layers are infinite-dimensional, we must assume that each space $\mathcal{X}_\ell$ is a reproducing kernel Hilbert space to reduce the problem to a finite-dimensional one, see (18). Assumption (11) is essential to ensure that the evaluation functional $f_1(f_2(x))$ at $x \in \mathcal{X}_{\ell-1}$ is jointly continuous in the pair $(f_1, f_2) \in \mathcal{X}_\ell \times \mathcal{X}_{\ell+1}$, as shown in Lemma A.8. This joint continuity, in turn, is crucial for establishing the existence of a minimizer of the regularized empirical risk minimization problem (9). We will see in Remark 5.4 and Remark 6.3 that such an assumption is easily satisfied by neural and discrete neural RKBS.

**Remark 5.1.** *In view of Assumption 1, since $\mathcal{X}_\ell$ is a reproducing kernel Hilbert space with a continuous reproducing kernel, $x(\cdot)$ is a continuous function, and hence $\rho_\ell$ is continuous on $\Theta_\ell$. Furthermore, given $x \in \mathcal{X}_\ell$, for all $\theta \in \Theta_\ell$ we have*

$$\begin{aligned}|\rho_\ell(x,\theta)| &\leq |\rho_\ell(0,\theta)| + |\rho_\ell(x,\theta) - \rho_\ell(0,\theta)| \\ &\leq C_\ell |\langle x, g_\ell(\theta)\rangle_{\mathcal{X}_\ell}| |\beta_\ell(\theta)| \\ &\leq C_\ell \|x\|_{\mathcal{X}_\ell} \sup_{\theta' \in \Theta_\ell} \|g_\ell(\theta')\|_{\mathcal{X}_\ell} |\beta_\ell(\theta)|\end{aligned}$$

*so that $\rho_\ell(x, \cdot) \in \mathcal{C}_0(\Theta_\ell)$. Thus, the integral RKBS $\mathcal{H}_\ell$, and therefore the associated deep integral RKBS, are well defined.*

**Remark 5.2.** *Recall that, for each $\ell = 1, \ldots, L$, the space $\Theta_\ell$ is second countable. Thus, for any $\mu \in \mathcal{M}(\Theta_\ell, \mathcal{X}_{\ell+1})$, its variation $|\mu|$ is a finite measure, so $|\mu|$ is regular, and therefore $\mu$ is regular too [12, Proposition 1]. Hence, taking into account that $\mathcal{X}_{\ell+1}$ is a Hilbert space, so that $\mathcal{X}'_{\ell+1} = \mathcal{X}_{\ell+1}$, by a generalization of the Riesz representation theorem the Banach space $\mathcal{M}(\Theta, \mathcal{X}_\ell)$ can be identified with the dual of $\mathcal{C}_0(\Theta_\ell, \mathcal{X}_{\ell+1})$, see [27, 25] for compact spaces, and [9] for second countable locally compact spaces. It follows that $\mathcal{M}(\Theta_\ell, \mathcal{X}_{\ell+1})$ can be endowed with the weak* topology, with respect to which the closed balls are compact.*

We can now derive the representer theorem for deep integral RKBS.

**Theorem 5.3** (Representer theorem for deep integral RKBS)**.** *Let $\mathcal{H}^{\text{deep}}$ be a deep integral RKBS of depth L, induced by a Banach space $\mathcal{H}$, from the input space $\mathcal{X}$ to the output space $\mathcal{Y} = \mathbb{R}^p$ satisfying Assumption 1. Assume that the loss function $\mathcal{L}$ is continuous in the first entry. Then, there exist $d_1, \ldots, d_{L+1} \in \mathbb{N}$ and, for all $\ell = 0, \ldots, L$,*

$$\begin{aligned}&\theta_1^{(\ell)}, \ldots, \theta_{d_\ell}^{(\ell)} \in \Theta_\ell, \\ &w_1^{(\ell+1)}, \ldots, w_{d_\ell}^{(\ell+1)} \in \widetilde{\mathcal{X}}_{\ell+1} \subset \mathcal{X}_{\ell+1} \quad \text{with} \quad \dim(\widetilde{\mathcal{X}}_{\ell+1}) \leq d_{\ell+1},\end{aligned}$$

*such that*

$$f^{\text{deep}}(x) = x^{(L+1)} \in \mathbb{R}^p, \qquad x \in \mathcal{X},$$

*with $x^{(L+1)}$ given by the recursive formula*

$$\begin{cases} x^{(0)} = x \in \mathcal{X} \\ x^{(\ell+1)} = \sum_{k=1}^{d_\ell} w_k^{(\ell+1)} \rho_\ell(x^{(\ell)}, \theta_k^{(\ell)}) \in \widetilde{\mathcal{X}}_{\ell+1}, \qquad \ell = 0, \ldots, L, \end{cases} \tag{12}$$

*is a solution of the minimization problem*

$$\min_{f^{\text{deep}} \in \mathcal{H}^{\text{deep}}} \mathcal{R}(f^{\text{deep}}) + \Phi(f^{\text{deep}}). \tag{13}$$

*Moreover, we have* $d_\ell \leq N d_{\ell+1}$ *for every* $\ell = 1. \dots, L$, *and*

$$\Phi(f^{\text{deep}}) \leq \sum_{\ell=0}^{L} \sum_{k=1}^{d_\ell} \|w_k^{(\ell+1)}\|_{\mathcal{X}_{\ell+1}}.$$

*Proof.* By definition of $\mathcal{H}^{\text{deep}}$ and the complexity measure $\Phi$, the minimization problem is equivalent to

$$\min_{f \in \mathcal{H}} \mathcal{R}(f^{\text{deep}}) + \|f\|_{\mathcal{H}},$$

where, for each $f = \oplus_{\ell=0}^{L} f_\ell \in \mathcal{H}$, the function $f^{\text{deep}} : \mathcal{X}_0 \to \mathcal{X}_{L+1}$ is the composition of $f_0, \dots, f_L$ ($\mathcal{X}_0 = \mathcal{X}$ and $\mathcal{X}_{L+1} = \mathbb{R}^p$). By construction, each $f_\ell$ is parameterized by some measure $\mu_\ell \in \mathcal{M}_\ell = \mathcal{M}(\Theta_\ell, \mathcal{X}_{\ell+1})$, according to (5). Moreover, by (6), the minimization problem (9) is equivalent to

$$\inf_{\mu \in \mathcal{M}} \mathcal{R}(f_\mu^{\text{deep}}) + \sum_{\ell=0}^{L} \|\mu_\ell\|_{\text{TV}} =: \inf_{\mu \in \mathcal{M}} \mathcal{S}(\mu), \tag{14}$$

where $\mathcal{M}$ is the Banach space $\bigoplus_{\ell=0}^{L} \mathcal{M}(\Theta_\ell, \mathcal{X}_{\ell+1})$ and, if $\mu = \mu_0 \oplus \cdots \oplus \mu_L \in \mathcal{M}$, $f_\mu^{\text{deep}}$ is the composition of $f_{\mu_0}, \dots, f_{\mu_L}$.

Fix a $\nu \in \mathcal{M}$, and let $R = \mathcal{S}(\nu)$. Then (14) is equivalent to

$$\inf_{\mu \in \prod_{\ell=0}^{L} B_{\mathcal{M}_\ell}(R)} \mathcal{S}(\mu). \tag{15}$$

Indeed, if $\mu$ is outside $\prod_{\ell=0}^{L} B_{\mathcal{M}_\ell}(R)$, then for some $\ell = 0, \dots, L$, $\|\mu_\ell\|_{\text{TV}} > R$, so that

$$\mathcal{S}(\mu) = \mathcal{R}(f_\mu) + \sum_{\ell=0}^{L} \|\mu_\ell\|_{\text{TV}} \geq \mathcal{R}(f_\mu) + \mathcal{S}(\nu) \geq \mathcal{S}(\nu),$$

which proves the equivalence. We now prove the existence of a minimizer of (15).

*Existence of a minimizer.* Assumption 1 along with Remark 5.1 ensures that, for any $\ell = 0, \dots, L$, the pair $(\rho_\ell, \rho_{\ell+1})$ of basis functions satisfies the conditions of Lemma A.8. Hence, taking into account Remark A.9, it follows that , for all $x \in \mathcal{X}$, the map

$$(\mu_0, \dots, \mu_L) \mapsto f_{\mu_L} \circ \cdots \circ f_{\mu_0}(x)$$

is *jointly* continuous from $\prod_{\ell=0}^{L} B_{\mathcal{M}_\ell}(R)$, endowed with the product topology induced by the weak$^*$ topology of each $B_{\mathcal{M}_\ell}(R)$, to $\mathcal{Y}$. Since $\mu_\ell \mapsto \|\mu\|_{\text{TV}}$ is weakly$^*$ continuous, the map $\mu \mapsto \mathcal{S}(\mu)$ is also continuous. Moreover, thanks to the Banach–Alaoglu theorem, the product $\prod_{\ell=0}^{L} B_{\mathcal{M}_\ell}(R)$ is weakly$^*$ compact. Hence, by the extreme value theorem, the problem (15) has at least a minimizer. Now we prove the representer theorem showing that a minimizer can be taken to have a finite width at every layer. Figure 1 summarizes the proof structure of the representer theorem.

*Representer theorem.* Let $\mu^*$ be any such solution. Let $x_i^{(0)} = x_i$ and $x_i^{(\ell+1)} = f_{\mu_\ell^*}(x_i^{(\ell)})$ for $\ell = 0, \dots, L$ and $i = 1, \dots, N$. Then, in view of Lemma A.4, a solution to (15) can be found by solving the following interpolation problems for all $\ell = 0, \dots, L$:

$$\inf_{\mu_\ell \in \mathcal{M}_\ell} \|\mu_\ell\|_{\text{TV}} \quad \text{subject to} \quad f_{\mu_\ell}(x_i^{(\ell)}) = x_i^{(\ell+1)} \quad i = 1, \dots, N. \tag{16}$$

*Case* $\ell = L$. Let us start from $\ell = L$. We want to apply Theorem A.2. To this end, let $\mathcal{U} = \mathcal{M}_L$ endowed with the weak* topology. We define $\mathcal{A} : \mathcal{U} \to \mathbb{R}^{N\times d_{L+1}}$ by

$$\mathcal{A}\mu = [f_{\mu_L}(x_i^{(L)})]_{i=1,\dots,N}.$$

Then $\mathcal{A}$ is a surjective continuous linear operator from $\mathcal{U}$ onto $H = \operatorname{Ran}\mathcal{A}$, with $\dim(H) \leq Nd_{L+1}$. Moreover, the norm $G = \|\cdot\|_{\mathrm{TV}}$ is coercive on $\mathcal{U}$. Indeed, by the Banach–Alaoglu theorem, the balls $B_{\mathcal{M}_L}(r)$ are weakly* compact for every $r > 0$. We define $F : H \to [0, \infty]$ by

$$F(h) = \begin{cases} 0 & h_i = x_i^{(L+1)} \text{ for all } i = 1,\dots,N \\ \infty & h_i \neq x_i^{(L+1)} \text{ for some } i = 1,\dots,N. \end{cases}$$

The function $F$ is the indicator function associated to the singleton $\{(x_1^{(L+1)},\dots,x_N^{(L+1)})\}$, so that it is convex, coercive and lower semi-continuous. Therefore, we can apply Theorem A.2 to derive that (16) has a solution of the form

$$\widetilde{\mu}_L = \sum_{k=1}^{d_L} c_k u_k,$$

for some $d_L \leq Nd_{L+1}$, $c_k > 0$ and $u_k \in \operatorname{Ext}(B_{\mathcal{M}_L}(1))$. By Lemma A.3, for each $k$ there are $y_k \in \operatorname{Ext}(B_{\mathcal{X}_{L+1}}(1))$ and $\theta_k^{(L)} \in \Theta_L$ such that

$$u_k = y_k \cdot \delta_{\theta_k^{(L)}}.$$

Thus, defining $w_{\cdot k}^{(L+1)} = c_k y_k \in \mathcal{X}_{L+1}$, we get

$$\widetilde{\mu}_L = \sum_{k=1}^{d_L} w_k^{(L+1)} \delta_{\theta_k^{(L)}}.$$

Thanks to Assumption 1, for all $x \in \mathcal{X}_L$ we have

$$x^{(L+1)} = f_{\widetilde{\mu}_L}(x) = \sum_{k=1}^{d_L} w_k^{(L+1)} \widetilde{\rho}_L(x(\theta_k^{(L)}), \theta_k^{(L)}). \tag{17}$$

Thus, (12) holds true for $\ell = L$ with $\widetilde{\mathcal{X}}_{L+1} = \mathcal{X}_{L+1} = \mathbb{R}^p$ and $d_{L+1} = p$.

*Case* $\ell = L - 1$. Now we consider $\ell = L - 1$. Let $K^{(L)}$ be the reproducing kernel of $\mathcal{X}_L$ and set

$$\widetilde{\mathcal{X}}_L = \operatorname{span}\{K^{(L)}(\cdot, \theta_k^{(L)}) : k = 1,\dots,d_L\} \subset \mathcal{X}_L, \qquad \dim \widetilde{\mathcal{X}}_L \leq d_L, \tag{18}$$

and $P_L$ the corresponding orthogonal projection from $\mathcal{X}_L$ onto $\widetilde{\mathcal{X}}_L$. By (17), it is clear that $f_{\widetilde{\mu}_L}(x) = f_{\widetilde{\mu}_L}(P_L x)$ for all $x \in \mathcal{X}_L$. Hence, since $\|P_L\mu\|_{\mathrm{TV}} \leq \|\mu\|_{\mathrm{TV}}$ for all $\mu \in \mathcal{M}(\Theta_{L-1}, \mathcal{X}_L)$, the direct sum of measures

$$(\mu^*)' = \mu_0^* \oplus \dots \oplus \mu_{L-2}^* \oplus P_L\mu_{L-1}^* \oplus \widetilde{\mu}_L$$

is a minimizer of (15). By Remark A.5, for each $i = 1,\dots,N$, we can replace $x_i^{(L)} = f_{\mu_{L-1}^*}(x_i^{(L-1)})$ with $P_L x_i^{(L)} = f_{P_L\mu_{L-1}^*}(x_i^{(L-1)})$, so that (16) for $\ell = L - 1$ reads as

$$\inf_{\mu_{L-1} \in \mathcal{M}(\Theta_{L-1}, \widetilde{\mathcal{X}}_L)} \|\mu_{L-1}\|_{\mathrm{TV}} \quad \text{subject to} \quad f_{\mu_{L-1}}(x_i^{(L-1)}) = P_L x_i^{(L)} \quad i = 1,\dots,N.$$

Since $\widetilde{\mathcal{X}}_L$ is finite dimensional, we can use again Theorem A.2, together with Lemma A.3, as in the previous step ($\ell = L$). In particular, this time we have $\dim(H) \leq N d_L$. Thus, we find a solution of the form

$$\widetilde{\mu}_{L-1} = \sum_{k=1}^{d_{L-1}} w_k^{(L)} \delta_{\theta_k^{(L-1)}},$$

for some $d_{L-1} \leq N d_L$ and $w_1^{(L)}, \dots, w_{d_{L-1}}^{(L)} \in \widetilde{\mathcal{X}}_L$. Now, for all $x \in \mathcal{X}_{L-1}$ we have

$$f_{\widetilde{\mu}_{L-1}}(x) = \sum_{k=1}^{d_{L-1}} w_k^{(L)} \widetilde{\rho}_{L-1}(x(\theta_k^{(L-1)}), \theta_k^{(L-1)}).$$

Iterating the argument, (12) holds true for $\ell = 1, \dots, L$.

*Case $\ell = 0$.* For the last step ($\ell = 0$), the minimization problem is

$$\inf_{\mu_0 \in \mathcal{M}_0} \|\mu_0\|_{\mathrm{TV}} \quad \text{subject to} \quad f_{\mu_0}(x_i) = P_1 x_i^{(1)} \in \widetilde{\mathcal{X}}_1 \quad i = 1, \dots, N,$$

which, as above, admits a solution $\widetilde{\mu}_0$ such that

$$f_{\widetilde{\mu}_0}(x) = \sum_{k=1}^{d_0} w_k^{(1)} \rho_0(x, \theta_k^{(0)}).$$

Then, $\widetilde{\mu} = \widetilde{\mu}_0 \oplus \dots \oplus \widetilde{\mu}_{L-1} \oplus \widetilde{\mu}_L$ is a solution of (15) and, consequently, $f_{\widetilde{\mu}}^{\mathrm{deep}} = f_{\widetilde{\mu}_0} \circ \dots \circ f_{\widetilde{\mu}_L}$ is a solution of the initial minimization problem (13). The bound on $d_\ell$ is also clear by iteration. For the bound on $\Phi(f^{\mathrm{deep}})$, by definition we have

$$\Phi(f^{\mathrm{deep}}) \leq \sum_{\ell=0}^{L} \|f_{\widetilde{\mu}_\ell}\|_{\mathcal{H}_\ell} \leq \sum_{\ell=0}^{L} \|\widetilde{\mu}_\ell\|_{\mathrm{TV}},$$

where, once again by Theorem A.2,

$$\|\widetilde{\mu}_\ell\|_{\mathrm{TV}} = \sum_{k=1}^{d_\ell} c_k.$$

But since $y_k \in \mathrm{Ext}(B_{\widetilde{\mathcal{X}}_{\ell+1}}(1))$, we have $\|y_k\|_{\widetilde{\mathcal{X}}_{\ell+1}} = 1$, whence, identifying again $\widetilde{\mathcal{X}}_{\ell+1}$ with a subspace of $\mathcal{X}_{\ell+1}$,

$$\sum_{k=1}^{d_\ell} c_k = \sum_{k=1}^{d_\ell} \|c_k y_k\|_{\mathcal{X}_{\ell+1}} = \sum_{k=1}^{d_\ell} \|w_k^{(\ell+1)}\|_{\mathcal{X}_{\ell+1}},$$

which concludes the proof. □

As established in Corollary 5.5, in the special case of a neural RKBS, the minimizer $f^{\mathrm{deep}}$ takes the form of a deep neural network with finite width at each hidden layer according to Definition 2.1.

**Remark 5.4.** *For neural RKBS, 10 is satisfied with $\widetilde{\rho}_\ell(t, n) = \sigma(t + c(\theta))\beta(\theta)$ for all $\ell = 1, \dots, L$. Furthermore, if $\sigma$ is Lipschitz with Lipschitz constant $C_\sigma$, 11 is satisfied with $C_\ell = C_\sigma$ and $g_\ell(\theta) = \delta_\theta$ for all $\ell = 1, \dots, L$. Indeed, for all $x, x' \in \mathcal{X}_\ell$ and $\theta \in \Theta_\ell$, we have that*

$$\begin{aligned} |\rho_\ell(x, \theta) - \rho_\ell(x', \theta)| &= |\sigma(x(\theta) + c_\ell(\theta)) - \sigma(x'(\theta) + c_\ell(\theta))||\beta_\ell(\theta)| \\ &\leq C_\sigma |x(\theta) - x'(\theta)||\beta_\ell(\theta)|. \end{aligned}$$

$$\begin{tikzcd}
\mathcal{X} \arrow[r, "f_{\mu_0^*}"] & \mathcal{X}_1 \arrow[r, "f_{\mu_1^*}"] & \cdots \arrow[r, "f_{\mu_{L-2}^*}"] & \mathcal{X}_{L-1} \arrow[r, "f_{\mu_{L-1}^*}"] & \mathcal{X}_L \arrow[r, "f_{\mu_L^*}"] & \mathbb{R}^p \\
\mathcal{X} \arrow[r, "f_{\mu_0^*}"] & \mathcal{X}_1 \arrow[r, "f_{\mu_1^*}"] & \cdots \arrow[r, "f_{\mu_{L-2}^*}"] & \mathcal{X}_{L-1} \arrow[r, "f_{P_L \mu_{L-1}^*}"] & \widetilde{\mathcal{X}}_L \arrow[r, "f_{\widetilde{\mu}_L}"] & \mathbb{R}^p \\
\mathcal{X} \arrow[r, "f_{\mu_0^*}"] & \mathcal{X}_1 \arrow[r, "f_{\mu_1^*}"] & \cdots \arrow[r, "f_{P_{L-1} \mu_{L-2}^*}"] & \widetilde{\mathcal{X}}_{L-1} \arrow[r, "f_{\widetilde{\mu}_{L-1}}"] & \widetilde{\mathcal{X}}_L \arrow[r, "f_{\widetilde{\mu}_L}"] & \mathbb{R}^p \\
 & & \vdots & & & \\
\mathcal{X} \arrow[r, "f_{\widetilde{\mu}_0}"] & \widetilde{\mathcal{X}}_1 \arrow[r, "f_{\widetilde{\mu}_1}"] & \cdots \arrow[r, "f_{\widetilde{\mu}_{L-2}}"] & \widetilde{\mathcal{X}}_{L-1} \arrow[r, "f_{\widetilde{\mu}_{L-1}}"] & \widetilde{\mathcal{X}}_L \arrow[r, "f_{\widetilde{\mu}_L}"] & \mathbb{R}^p
\end{tikzcd}$$

FIGURE 1. Proof structure of Theorem 5.3.

**Corollary 5.5** (Representer theorem for neural RKBS). *Under the assumptions of Theorem 5.3, if $\mathcal{H}^{\text{deep}}$ is a neural RKBS, then there exist*

$$d_1, \dots, d_L \in \mathbb{N}, \qquad d_\ell \leq N d_{\ell+1},$$
$$W^{(\ell)} \in \mathbb{R}^{d_\ell \times d_{\ell-1}}, \qquad b^{(\ell)} \in \mathbb{R}^{d_\ell}, \qquad \ell = 1, \dots, L+1,$$

*such that*

$$f^{\text{deep}}(x) = x^{(L+1)} \in \mathbb{R}^p, \qquad x \in \mathcal{X}, \tag{19}$$

*with $x^{(L+1)}$ given by recursive formula*

$$\begin{cases} x^{(1)} = W^{(1)} x + b^{(1)} \\ x^{(\ell+1)} = W^{(\ell+1)} \sigma(x^{(\ell)}) + b^{(\ell+1)}, \qquad \ell = 1, \dots, L, \end{cases} \tag{20}$$

*is a solution of the minimization problem*

$$\min_{f^{\text{deep}} \in \mathcal{H}^{\text{deep}}} \mathcal{R}(f^{\text{deep}}) + \Phi(f^{\text{deep}}). \tag{21}$$

*Proof.* Revisiting the proof of Theorem 5.3, for all $\ell = 1, \dots, L+1$, we can choose a basis $\{e_k^{(\ell)}\}_{k=1}^{\dim \mathcal{X}_\ell}$ of $\mathcal{X}_\ell$ such that

$$\widetilde{\mathcal{X}}_\ell \subset \operatorname{span}\{e_1^{(\ell)}, \dots, e_{d_\ell}^{(\ell)}\},$$

so that the elements $w_1^{(\ell)}, \dots, w_{d_{\ell-1}}^{(\ell)} \in \widetilde{\mathcal{X}}_\ell$ can be identified with vectors in $\mathbb{R}^{d_\ell}$ and collected in a $d_\ell \times d_{\ell-1}$ matrix

$$U^{(\ell)} = \left( \, w_1^{(\ell)} \beta_{\ell-1}(\theta_1^{(\ell-1)}) \, \middle| \, \cdots \, \middle| \, w_{d_{\ell-1}}^{(\ell)} \beta_{\ell-1}(\theta_{d_{\ell-1}}^{(\ell-1)}) \, \right),$$

where $\beta_0 = 1$. Similarly, for all $\ell = 1, \dots, L$, the elements $K^{(\ell)}(\cdot, \theta_1^{(\ell)}), \dots, K^{(\ell)}(\cdot, \theta_{d_\ell}^{(\ell)}) \in \widetilde{\mathcal{X}}_\ell \simeq \mathbb{R}^{d_\ell}$ define a $d_\ell \times d_\ell$ matrix

$$V^{(\ell)} = \left( \, K^{(\ell)}(\cdot, \theta_1^{(\ell)})^\top \, \middle| \, \cdots \, \middle| \, K^{(\ell)}(\cdot, \theta_{d_\ell}^{(\ell)})^\top \, \right),$$

and the offsets $c_\ell(\theta_1^{(\ell)}), \dots, c_\ell(\theta_{d_\ell}^{(\ell)}) \in \mathbb{R}$ a vector

$$c^{(\ell)} = \left( c_\ell(\theta_1^{(\ell)}), \dots, c_\ell(\theta_{d_\ell}^{(\ell)}) \right)^\top \in \mathbb{R}^{d_\ell}.$$

Hence, (12) reads as

$$\begin{cases} x^{(0)} = (x(\theta_1^{(0)}), \dots, x(\theta_{d_0}^{(0)}))^\top \\ x^{(1)} = U^{(1)} x^{(0)} \\ x^{(\ell+1)} = U^{(\ell+1)} \left( \sigma(V^{(\ell)} x^{(\ell)} + c^{(\ell)}) \right), \qquad \ell = 1, \dots, L. \end{cases}$$

For all $\ell = 1, \dots, L+1$, define the matrix

$$W^{(\ell)} = V^{(\ell)} U^{(\ell)} \in \mathbb{R}^{d_\ell \times d_{\ell-1}},$$

where $V^{(L+1)} = \mathrm{Id}_{d_{L+1} \times d_{L+1}}$, and the vector

$$b^{(\ell)} = c^{(\ell)} \in \mathbb{R}^{d_\ell},$$

where $c^{(L+1)} = 0 \in \mathbb{R}^p$. Up to redefining the points $x^{(\ell)}$ for $\ell = 1, \dots, L$, a solution $f^{\text{deep}}$ of the empirical risk minimization problem (21) is given by $f^{\text{deep}}(x) = x^{(L+1)} \in \mathbb{R}^p$ with $x^{(L+1)}$ given by the recursive formula

$$\begin{cases} x^{(0)} = (x(\theta_1^{(0)}), \dots, x(\theta_{d_0}^{(0)}))^\top \\ x^{(1)} = W^{(1)} x^{(0)} + b^{(1)} \\ x^{(\ell+1)} = W^{(\ell+1)} \sigma(x^{(\ell)}) + b^{(\ell+1)}, \qquad \ell = 1, \dots, L, \end{cases}$$

which concludes the proof. □

## 6. Discrete Neural RKBS

Next, we construct a particular instance of neural RKBS with a countably infinite number of neurons per hidden layer, which we call *discrete* neural RKBS, see Figure 2. This further specialization to these spaces allows for a more explicit characterization of the complexity measure $\Phi$ in the corresponding representer theorem.

In the following formulation, differently from the one in Remark 4.4, the spaces $\Theta_\ell$ of the hidden layers play the role of index spaces for the parameters, and they do not directly correspond to the spaces where parameters live.

**Definition 6.1** (Discrete Neural RKBS). *Fix the parameter spaces $\Theta_\ell$ and the layer spaces $\mathcal{X}_\ell$ as follows*

$$\begin{aligned} &\Theta_0 = \{0, \dots, d\}, && \mathcal{X}_0 = \ell^2(\{1, \dots, d\}) = \mathbb{R}^d, && \\ &\Theta_\ell = \mathbb{N}, && \mathcal{X}_\ell = \ell^2(\mathbb{N}), && \ell = 1, \dots, L, \\ &\Theta_{L+1} = \{1, \dots, p\}, && \mathcal{X}_{L+1} = \ell^2(\{1, \dots, p\}) = \mathbb{R}^p, && \ell = L+1, \end{aligned}$$

*and let $\sigma : \mathbb{R} \to \mathbb{R}$ be a Lipschitz activation function such that $\sigma(0) = 0$. Then, set*

$$\rho_0(x, n) = \begin{cases} 1 & n = 0 \\ x_n & n = 1, \dots, d \end{cases}, \qquad x \in \mathbb{R}^d$$

$$\rho_\ell(x, n) = \begin{cases} 1 & n = 0 \\ \sigma(x_{n-1}) & n \geq 1 \end{cases}, \qquad x \in \ell^2(\mathbb{N}), \qquad \ell = 1, \dots, L\,.$$

*We call the resulting space $\mathcal{H}^{\text{deep}}$ a* discrete *neural RKBS.*

$$\begin{tikzcd}
\mathbb{R}^d \arrow[r, "f_0"] \arrow[rrrr, bend right, "f^{\mathrm{deep}}"'] & \ell^2(\mathbb{N}) \arrow[r] & \cdots \arrow[r] & \ell^2(\mathbb{N}) \arrow[r, "f_L"] & \mathbb{R}^p
\end{tikzcd}$$

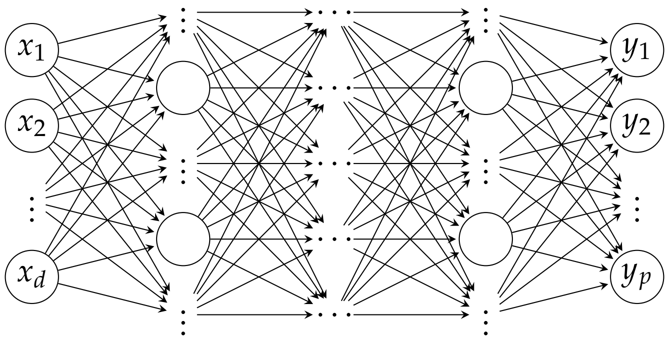

FIGURE 2. Architecture of discrete neural RKBS functions.

**Remark 6.2.** *Some observations on Definition 6.1 are in order. First, note that $\mathcal{X}_0$ can be thought of as a function space on $\Theta_0$ by putting $x(0) = 1$ for all $x \in \mathcal{X}_0$. Second, the Lipschitzianity of $\sigma$ implies that*

$$\sigma(x) \in \ell^2(\mathbb{N}), \qquad \text{for all } x \in \ell^2(\mathbb{N}), \tag{22}$$

*where $\sigma$ applies on sequences component by component. Indeed, if $C_\sigma$ denotes the Lipschitz constant of $\sigma$, we have*

$$|\sigma(x_n)| = |\sigma(x_n) - \sigma(0)| \leq C_\sigma |x_n - 0| = C_\sigma |x_n|.$$

*In particular, this implies that*

$$\lim_{n\to\infty} \sigma(x_n) = 0, \qquad \text{for all } x \in \ell^2(\mathbb{N}) \,. \tag{23}$$

*Finally, assuming $\sigma(0) = 0$ is not restrictive. In fact, if $\sigma(0) \neq 0$, the choice of the activation function $\sigma' = \sigma - \sigma(0)$ simply implies a scaling of the offsets.*

**General form of discrete neural RKBS functions.** An element $f^{\mathrm{deep}}$ of the neural RKBS $\mathcal{H}^{\mathrm{deep}}$ is a composition of $L+1$ integral functions $f_0,\ldots,f_L$, where each $f_\ell$ is defined via a measure $\mu_\ell$ by $f_\ell = f_{\mu_\ell}$. We now derive an explicit expression for elements in a discrete neural RKBS, showing that these functions correspond to infinite-width neural networks with a countably infinite number of neurons per hidden layer, see Figure 2.

*Case $\ell = 0$.* From layer 0 to layer 1, we have

$$\mu_0 \in \mathcal{M}(\{0,\ldots,d\}, \ell^2(\mathbb{N})),$$

so that

$$\mu_0 = \sum_{m=0}^{d} w_m^{(1)} \delta_m,$$

for a family of $d+1$ vectors $w_0^{(1)}, \ldots w_d^{(1)} \in \ell^2(\mathbb{N})$. Let $b^{(1)} = w_0^{(1)} \in \ell^2(\mathbb{N})$ and define the bounded operator

$$W^{(1)} \in B(\mathbb{R}^d, \ell^2(\mathbb{N})), \qquad W^{(1)}x = \sum_{m=1}^{d} w_m^{(1)} x_m.$$

Then, for $x \in \mathbb{R}^d$, the function $f_0 : \mathbb{R}^d \to \ell^2(\mathbb{N})$ is

$$f_0(x) = \sum_{m=0}^{d} w_m^{(1)} \rho_0(x,m) = w_0^{(1)} + \sum_{m=1}^{d} w_m^{(1)} x_m = W^{(1)}x + b^{(1)} \in \ell^2(\mathbb{N}),$$

and the scalar components of $f_0$ are

$$f_0(x)_n = \langle x, w_{n\cdot}^{(1)} \rangle_{\mathbb{R}^d} + b_n^{(1)}, \qquad n \in \mathbb{N},$$

where $w_{n\cdot}^{(1)} \in \mathbb{R}^d$ with $w_{nm}^{(1)} = (w_m^{(1)})_n$.

*Case* $\ell = 1, \ldots, L-1$. From layer $\ell$ to layer $\ell + 1$ we have

$$\mu_\ell \in \mathcal{M}(\mathbb{N}, \ell^2(\mathbb{N})),$$

so that

$$\mu_\ell = \sum_{m=0}^{\infty} w_m^{(\ell+1)} \delta_m,$$

for a countable family of vectors $w_0^{(\ell+1)}, \ldots, w_m^{(\ell+1)}, \ldots \in \ell^2(\mathbb{N})$ such that

$$\|\mu_\ell\|_{\mathrm{TV}} = \sum_{m=0}^{\infty} \|w_m^{(\ell+1)}\|_{\ell^2(\mathbb{N})} < \infty. \tag{24}$$

As before, set $b^{(\ell+1)} = w_0^{(\ell+1)} \in \ell^2(\mathbb{N})$ and

$$W^{(\ell+1)} \in B(\ell^2(\mathbb{N}), \ell^2(\mathbb{N})), \qquad W^{(\ell+1)}x = \sum_{m=1}^{+\infty} w_m^{(\ell+1)} x_{m-1},$$

where the series converges absolutely in $\ell_2(\mathbb{N})$ due to (24). Hence, for $x \in \ell_2(\mathbb{N})$,

$$\begin{aligned} f_\ell(x) &= \sum_{m=0}^{+\infty} w_m^{(\ell+1)} \rho_\ell(x,m) = w_0^{(\ell+1)} + \sum_{m=1}^{+\infty} w_m^{(\ell+1)} \sigma(x_{m-1}) \\ &= W^{(\ell+1)}\left(\sigma(x)\right) + b^{(\ell+1)} \in \ell^2(\mathbb{N}), \end{aligned}$$

where $\sigma(x) \in \ell^2(\mathbb{N})$ by (22). The component of $f_\ell(x)$ are

$$f_\ell(x)_n = \langle \sigma(x), w_{n\cdot}^{(\ell+1)} \rangle_{\ell^2} + b_n^{(\ell+1)}, \qquad n \in \mathbb{N},$$

where $w_{n\cdot}^{(\ell+1)} \in \ell^2(\mathbb{N})$ and $w_{nm}^{(\ell+1)} = (w_m^{(\ell+1)})_n$.

*Case* $\ell = L$. Finally, from layer $L$ to layer $L+1$, we have

$$\mu_L \in \mathcal{M}(\mathbb{N}, \mathbb{R}^p),$$

so that

$$\mu_L = \sum_{m=0}^{\infty} w_m^{(L+1)} \delta_m,$$

for a countable family of vectors $w_0^{(L+1)}, \ldots, w_m^{(L+1)}, \ldots \in \mathbb{R}^p$ such that

$$\sum_{m=0}^{\infty} \|w_m^{(\ell+1)}\|_{\mathbb{R}^p} < +\infty. \tag{25}$$

As before, set $b^{(L+1)} = w_0^{(L+1)} \in \ell^2(\mathbb{N})$ and

$$W^{(L+1)} \in B(\ell^2(\mathbb{N}), \mathbb{R}^p), \qquad W^{(L+1)}x = \sum_{m=1}^{+\infty} w_m^{(L+1)} x_{m-1},$$

where the series converges absolutely in $\mathbb{R}^p$ due to (25). Hence, for $x \in \ell_2(\mathbb{N})$,

$$\begin{aligned} f_L(x) &= \sum_{m=0}^{+\infty} w_m^{(L+1)} \rho_L(x,m) = w_0^{(L+1)} + \sum_{m=1}^{+\infty} w_m^{(L+1)} \sigma(x_{m-1}) \\ &= W^{(L+1)}\left(\sigma(x)\right) + b^{(L+1)} \in \ell^2(\mathbb{N}), \end{aligned}$$

where $\sigma(x) \in \ell^2(\mathbb{N})$ by (22). The component of $f_L(x)$ are

$$f_L(x)_n = \langle \sigma(x), w_{n\cdot}^{(L+1)} \rangle_{\ell^2} + b_n^{(L+1)}, \qquad n = 1, \ldots, p, \tag{26}$$

where $w_{n\cdot}^{(L+1)} \in \ell^2(\mathbb{N})$ and $w_{nm}^{(L+1)} = (w_m^{(L+1)})_n$.

By iteration $x^{(\ell+1)} = f_\ell(x^{(\ell)})$, we obtain

$$\begin{cases} x^{(0)} = x & \in \mathbb{R}^d \\ x^{(1)} = W^{(1)} x^{(0)} + b^{(1)} & \in \ell^2(\mathbb{N}) \\ x^{(\ell+1)} = W^{(\ell+1)}\left(\sigma(x^{(\ell)})\right) + b^{(\ell+1)} & \in \ell^2(\mathbb{N}) \\ x^{(L+1)} = W^{(L+1)}\left(\sigma(x^{(L)})\right) + b^{(L+1)} & \in \mathbb{R}^p\,. \end{cases}$$

We stress that the width of the $L$ hidden layers $\ell = 1, \ldots, L$ is infinite and countable (the neurons are parameterized by $\mathbb{N}$), while input and output layers $\ell = 0$ and $\ell = L+1$ have fixed finite widths $d$ and $p$, respectively. Also note that infinite-width neural networks generalize finite-width neural networks, where the inner layers are generated by infinite-rank operators.

We can visualize the shallow ($L = 1$), and the simplest non-shallow case ($L = 2$), considering a non-iterative expression. For $L = 1$, we have

$$f(x) = W^{(2)}\left(\sigma(W^{(1)}x + b^{(1)})\right) + b^{(2)}\,.$$

In the case of $L = 2$ hidden layers, we can write

$$f(x) = W^{(3)}\left(\sigma\left(W^{(2)}\left(\sigma(W^{(1)}x + b^{(1)})\right) + b^{(2)}\right)\right) + b^{(3)}\,.$$

Comparing our construction to the one proposed in [22], we remark that our networks do not have any rank constraint, in the sense that every hidden layer has infinte width.

**Finite form of discrete neural RKBS functions.** We now show that the neural functions defined in 2.1 correspond to measures $\mu_1, \ldots, \mu_L$ having finite support. Indeed, under this assumption, for each $\ell = 1, \ldots, L$,

$$\mu_\ell = b^{(\ell+1)}\delta_0 + \sum_{k=1}^{d_\ell} w_k^{(\ell+1)} \delta_{m_k^{(\ell)}},$$

for some $m_1^{(\ell)}, \ldots, m_{d_\ell}^{(\ell)} \in \mathbb{N} \setminus \{0\}$ and some $w_1^{(\ell+1)}, \ldots, w_k^{(\ell+1)}$ that are in $\ell_2(\mathbb{N})$ if $\ell < L$, and in $\mathbb{R}^p$ if $\ell = L$.

We define $f_\ell$ starting from the last layer. Since the support of $\mu_L$ is $\{0, m_1^{(L)}, \ldots, m_{d_L}^{(L)}\}$, by (26) $f_L$ depends only on the variables $m_1^{(L)}, \ldots, m_{d_L}^{(L)}$, so that we can regard $f_L$ as a function from $\mathbb{R}^{d_L}$ to $\mathbb{R}^p$ given by

$$f_L(x) = W^{(L+1)}\sigma(x) + b,$$

where $W^{(L+1)}$ is the $d_L \times d_{L+1}$ matrix (recall that $d_{L+1} = p$) with components

$$W^{(L+1)}_{nk} = (w^{(L+1)}_{m^{(L)}_k})_{n-1}, \qquad n = 1, \dots, p, \qquad k = 1, \dots, d_L.$$

Since $f_L$ is defined on $\mathbb{R}^{d_L}$, regarded as a finite-dimensional subspace of $\ell_2(\mathbb{N})$, denoting by $P : \ell_2(\mathbb{N}) \to \mathbb{R}^{d_L}$ the corresponding projection

$$Px = (x_{m^{(L)}_1}, \dots, x_{m^{(L)}_{d_L}}),$$

for all $x \in \ell_2(\mathbb{N})$ we have

$$f_L(f_{L-1}(x)) = f_L(f_{\mu_{L-1}}(x)) = f_L(f_{P\mu_{L-1}}(x)).$$

Then, without loss of generality, we can assume that the measure $\mu_{L-1}$ is in $\mathcal{M}(\mathbb{N}, \mathbb{R}^{d_L})$ and it has a finite support. By iterating this procedure, we can assume that for all $\ell$

$$\mu_\ell \in \mathcal{M}(\{0, \dots, d_\ell\}, \mathbb{R}^{d_{\ell+1}}),$$

for some $d_0, d_1, \dots, d_L, d_{L+1} \in \mathbb{N}$ (with $d_0 = d$ and $d_{L+1} = p$). This means that

$$\mu_\ell = b^{(\ell+1)}\delta_0 + \sum_{k=1}^{d_\ell} w^{(\ell+1)}_k \delta_k.$$

For all $\ell = 1, \dots, L+1$, let $W^{(\ell)}$ be the $n_{d_{\ell-1}} \times n_{d_\ell}$ matrix

$$W^{(\ell)}_{nk} = (w^{(\ell)}_k)_n \qquad n = 1, \dots, d_\ell,\ k = 1, \dots, d_{\ell-1}.$$

Then $f_\ell = f_{\mu_\ell} : \mathbb{R}^{d_{\ell-1}} \to \mathbb{R}^{d_\ell}$ is given by

$$f_\ell(x) = \begin{cases} W^{(1)}x + b^{(1)} & \ell = 0 \\ W^{(\ell+1)}\sigma(x) + b^{(\ell+1)} & \ell > 1 \end{cases},$$

$$f^{\text{deep}} = f_L \circ \dots \circ f_0,$$

is a neural deep function according to Definition 2.1, and we can rewrite $f^{\text{deep}}$ as

$$\begin{cases} x^{(1)} = W^{(1)}x + b^{(1)} & \in \mathbb{R}^{d_1} \\ x^{(\ell+1)} = W^{(\ell+1)}\sigma(x^{(\ell)}) + b^{(\ell+1)} & \in \mathbb{R}^{d_{\ell+1}} \qquad \ell = 1, \dots, L, \end{cases}$$

that is, $f^{\text{deep}}$ is a neural network of depth $L$ and (finite) widths $d_1, \dots, d_L$.

**Remark 6.3.** *For discrete neural RKBS, Equation* (10) *is satisfied replacing the basis functions in Definition 6.1 by $\rho_\ell(x, n) = \sigma(x_{n-1})\beta_{n-1}$ for all $\ell = 1, \dots, L$, with $\sigma \colon \mathbb{R} \to \mathbb{R}$ Lipschitz and $\beta : \mathbb{N} \to \mathbb{R}$ a positive sequence converging to zero. Furthermore, Equation* (11) *is satisfied with $C_\ell = C_\sigma$ and $g_\ell(n) = \delta_{n-1}$ for all $\ell = 1, \dots, L$. Indeed, for all $x, x' \in \ell^2(\mathbb{N})$ and $n \in \mathbb{N}$,*

$$|\rho_\ell(x, n) - \rho_\ell(x', n)| \le |\sigma(x_{n-1}) - \sigma(x'_{n-1})||\beta_{n-1}| \le C_\sigma |x_{n-1} - x'_{n-1}||\beta_{n-1}|.$$

**Corollary 6.4** (Representer theorem for discrete neural RKBS)**.** *Under the assumptions of Theorem 5.3, if $\mathcal{H}^{\text{deep}}$ is a discrete neural RKBS, then the claim of Corollary 5.5 holds true. Moreover,*

$$\Phi(f^{\text{deep}}) \le \sum_{\ell=0}^{L} \sum_{k=1}^{d_\ell} \left( \sum_{j=1}^{d_{\ell+1}} |W^{(\ell+1)}_{jk} \beta_k^{-1}|^2 \right)^{1/2}.$$

*Proof.* Since $\Theta_0 = \{0, \ldots, d\}$, without loss of generality we can assume $d_0 = d + 1$ and $\theta_n^{(0)} = n$ for all $k = 0, \ldots, d_0$. Taking into account that $\rho_0(x, 0) = 1$ and $\rho_0(x, k) = x_k$ if $1 \leq k \leq d$, we have

$$x^{(0)} = (1, x) \in \mathbb{R} \times \mathbb{R}^d.$$

Since, for $\ell = 1, \ldots, L$, $\Theta_\ell = \mathbb{N}$ and $K_\ell(\cdot, n) = e_{n-1}$ where $\{e_n\}_{n \in \mathbb{N}}$ is the canonical base of $\mathcal{X}_\ell = \ell_2(\mathbb{N})$, up to a permutation, $V^{(\ell)}$ is the identity and the claim follows. □

The above result shows that deep neural networks with finite width at each hidden layer are optimal, in the sense that they are solutions of empirical risk minimization over neural RKBS. Moreover, it provides an upper bound on the network width depending on sample size and input/output dimensions. Finally, it shows that the regularization norm is controlled by the $\ell^1$ norm of the $\ell^2$ norms of the weights of the network.

## 7. Conclusions and future work

Studying function spaces defined by neural networks provides a natural way to understand their properties. Recently, reproducing kernel Banach spaces have emerged has a useful concept to study shallow networks.

In this paper, we take a step towards more complex architectures considering deep networks. We allow for a wide class of activation functions and remove unnecessary low rank constraints. Our main contributions are defining classes of neural RKBS obtained composing vector-valued RKBS and deriving corresponding representer theorems borrowing ideas from [22].

Future developments include considering more structured architectures, for example convolutional networks, as well as investigating the statistical and computational properties of neural RKBS. Moreover, finer characterizations of the Banach structure could be obtained using the specific form of the activation function and the functional properties that this induces (see [22] for the ReLU).

## Acknowledgments

L. R. acknowledges the financial support of the European Research Council (grant SLING 819789), the European Commission (Horizon Europe grant ELIAS 101120237), the US Air Force Office of Scientific Research (FA8655-22-1-7034), the Ministry of Education, University and Research (FARE grant ML4IP R205T7J2KP; grant BAC FAIR PE00000013 funded by the EU - NGEU) and the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. This work represents only the view of the authors. The European Commission and the other organizations are not responsible for any use that may be made of the information it contains. The research by E.D.V. and L. R. has been supported by the MIUR grant PRIN 202244A7YL. The research by E.D.V. has been supported by the MIUR Excellence Department Project awarded to Dipartimento di Matematica, Università di Genova, CUP D33C23001110001. 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). This work represents only the view of the authors. The European Commission and the other organizations are not responsible for any use that may be made of the information it contains. This work was partially supported by the MUR Excellence Department Project MatMod@TOV awarded to the Department of Mathematics, University of Rome Tor Vergata, CUP E83C18000100006. S.V. also acknowledges financial support from the MUR 2022 PRIN project GRAFIA, project code 202284Z9E4.

## References

[1] N. Aronszajn. "Theory of Reproducing Kernels". In: *Transactions of the American Mathematical Society* 68.3 (1950), pp. 337–404.

[2] F. Bach. "Breaking the Curse of Dimensionality with Convex Neural Networks". In: *Journal of Machine Learning Research* 18.19 (2017), pp. 1–53.

[3] F. Bartolucci, E. De Vito, L. Rosasco, and S. Vigogna. "Understanding neural networks with reproducing kernel Banach spaces". In: *Applied and Computational Harmonic Analysis* 62 (2023), pp. 194–236.

[4] A. Bietti and F. Bach. "Deep equals shallow for ReLu networks in kernel regimes". In: *International Conference on Learning Representations (ICLR)* 9 (2021).

[5] A. Bietti and J. Mairal. "On the Inductive Bias of Neural Tangent Kernels". In: *Advances in Neural Information Processing Systems (NeurIPS)* 32 (2019).

[6] V. I. Bogachev. *Measure theory*. Vol. II. Springer-Verlag, 2007.

[7] K. Bredies and M. Carioni. "Sparsity of solutions for variational inverse problems with finite-dimensional data". In: *Calculus of Variations and Partial Differential Equations* 59.14 (2020).

[8] C. Carmeli, E. De Vito, and A. Toigo. "Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem". In: *Analysis and Applications* 4.4 (2006), pp. 377–408.

[9] C. Carmeli, E. De Vito, A. Toigo, and V. Umanità. "Vector valued reproducing kernel Hilbert spaces and universality". In: *Analysis and Applications* 8.01 (2010), pp. 19–61.

[10] L. Chizat, E. Oyallon, and F. Bach. "On lazy training in differentiable programming". In: *Advances in Neural Information Processing Systems (NeurIPS)* 32 (2019).

[11] J. Diestel and J. Uhl. *Vector Measures*. American Mathematical Society, 1977.

[12] N. Dinculeanu. "Sur la représentation intégrale des certaines opérations linéaires. III". In: *Proc. Amer. Math. Soc.* . 10 (1959), pp. 59–68.

[13] N. Dinculeanu. *Vector integration and stochastic integration in Banach spaces*. John Wiley & Sons, 2000.

[14] N. Dunford and J. T. Schwartz. *Linear Operators: With the Assistance of William G. Bade and Robert G. Bartle*. Interscience., 1963.

[15] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. "When Do Neural Networks Outperform Kernel Methods?" In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. 2020, pp. 14820–14830.

[16] B. Hanin. "Random Neural Networks in the Infinite Width Limit as Gaussian Processes". In: *Annals of Applied Probability* to appear (2023).

[17] A. Jacot, C. Hongler, and F. Gabriel. "Neural Tangent Kernel: Convergence and Generalization in Neural Networks". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018, pp. 8580–8589.

[18] R. R. Lin, H. Z. Zhang, and J. Zhang. "On Reproducing Kernel Banach Spaces: Generic Definitions and Unified Framework of Constructions". In: *Acta Mathematica Sinica* 38.8 (2022), pp. 1459–1483.

[19] R. M. Neal. *Bayesian Learning for Neural Networks*. Vol. 118. Springer, 2012.

[20] G. Ongie, R. Willett, D. Soudry, and N. Srebro. "A Function Space View of Bounded Norm Infinite Width ReLU Nets: The Multivariate Case". In: *Eighth International Conference on Learning Representations (ICLR)*. 2020.

[21] R. Parhi and R. D. Nowak. "Banach Space Representer Theorems for Neural Networks and Ridge Splines". In: *Journal of Machine Learning Research* 22.43 (2021), pp. 1–40.

[22] R. Parhi and R. D. Nowak. "What Kinds of Functions Do Deep Neural Networks Learn? Insights from Variational Spline Theory". In: *SIAM Journal on Mathematics of Data Science* 4.2 (2022), pp. 464–489.

[23] A. Rahimi and B. Recht. "Random Features for Large-Scale Kernel Machines". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 20. 2007.

[24] W. Rudin. *Functional Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, New York, 1991.

[25] R. Ryan. "The F. and M. Riesz theorem for vector measures". In: *Indag. Math.* 25 (1963), pp. 408–412.

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

[27] I. Singer. "Linear functionals on the space of continuous mappings of a compact Hausdorff space into a Banach spaces". In: *Rev. Math. Pures Appl.* 2 (1957), pp. 301–315.

[28] M. Unser. "A Unifying Representer Theorem for Inverse Problems and Machine Learning". In: *Foundations of Computational Mathematics* (2020), pp. 1–20.

[29] M. Unser. "From kernel methods to neural networks: A unifying variational formulation". In: *Foundations of Computational Mathematics* (2023), pp. 1–40.

[30] M. Unser. "Ridges, Neural Networks, and the Radon Transform". In: *Journal of Machine Learning Research* 24.37 (2023), pp. 1–33.

[31] M. Unser and J. Fageot. "Native Banach spaces for splines and variational inverse problems". In: *arXiv:1904.10818* (2019).

[32] M. Unser, J. Fageot, and J. P. Ward. "Splines are universal solutions of linear inverse problems with generalized TV regularization". In: *SIAM Review* 59.4 (2017), pp. 769–793.

[33] D. Werner. "Extreme points in spaces of operators and vector–valued measures". In: *Proceedings of the 12th Winter School on Abstract Analysis* (1984), pp. 135–143.

[34] H. Zhang, Y. Xu, and J. Zhang. "Reproducing Kernel Banach Spaces for Machine Learning". In: *Journal of Machine Learning Research* 10.95 (2009), pp. 2741–2775.

## Appendix A. Sparse solutions to finite-dimensional variational problems

The key ingredient to establish our representer theorem is given by a powerful variational result proved in [7]. This result deals with general minimization problems with finite-dimensional constraints and seminorm penalization. It states that such problems admit sparse solutions, namely finite linear combinations of extremal points of the seminorm unit ball. We report the formal statement below, after recalling the definition of extremal point.

**Definition A.1** (Extremal point). *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, that is, there do not exist $p, r \in Q$, $p \neq r$, such that $q = tp + (1-t)r$ for some $t \in (0,1)$. We denote the set of extremal points of $Q$ by* $\mathrm{Ext}(Q)$.

**Theorem A.2** ([7, Theorem 3.3]). *Consider the problem*

$$\arg\min_{u \in U} F(\mathcal{A}u) + G(u), \tag{27}$$

*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. Let $B_U(1)$ denote the unit ball $\{u \in U : G(u) \leq 1\}$. Then* (27) *has solutions of the form*

$$\sum_{k=1}^{K} c_k u_k,$$

*with $u_k \in \mathrm{Ext}(B_U(1))$, $c_k > 0$, $K \leq \dim H$, and $\sum_{k=1}^{K} c_k = G(u)$.*

Theorem A.2 is a simplified version of [7, Theorem 3.3], where $G$ is only assumed to be a seminorm.

In view of Theorem A.2, in order to determine the form of sparse solutions we need to characterize the set of extremal point of the unit ball in $U$. In our construction of integral RKBS, $U$ is the space of vector measures with total variation norm. The following result provides the desired characterization for our case. It appears in [33] assuming that $\Theta$ is compact, but the proof works analogously when $\Theta$ is locally compact. We report the proof for the reader's convenience.

**Lemma A.3** ([33, Theorem 2]). *Let $\Theta$ be a locally compact, second countable topological space, and let $\mathcal{Y}$ be a Banach space. Then*

$$\mathrm{Ext}(B_{\mathcal{M}(\Theta,\mathcal{Y})}(1)) = \{y \cdot \delta_\theta : y \in \mathrm{Ext}(B_{\mathcal{Y}}(1)), \theta \in \Theta\}.$$

*Proof.* Let us denote $E = \{y \cdot \delta_\theta : y \in \mathrm{Ext}(B_{\mathcal{Y}}(1)), \theta \in \Theta\}$. We start showing that $\mathrm{Ext}(B_{\mathcal{M}}(1)) \subseteq E$. Suppose that $\mu \in \mathrm{Ext}(B_{\mathcal{M}}(1))$ but $\mu \neq y \cdot \delta_\theta$ for any $y \in \mathrm{Ext}(B_{\mathcal{Y}}(1))$ and $\theta \in \Theta$. Then $|\mu| \neq \delta_\theta$ for any $\theta \in \Theta$, and there is $A \in \mathcal{B}(\Theta)$ such that $0 < |\mu|(A) < 1$. Denote by $\chi_A$ the indicator function on $A$. Then, setting $t = |\mu|(A)$, $\mu_1 = \mu\chi_A / t$ and $\mu_2 = \mu\chi_{\Theta\setminus A}/(1-t)$, we can write $\mu$ as a convex combination

$$\mu = t\mu_1 + (1-t)\mu_2. \tag{28}$$

Since $t \in (0,1)$ and $\mu_1, \mu_2 \in B_{\mathcal{M}}(1)$, we get that $\mu \notin \mathrm{Ext}(B_{\mathcal{M}}(1))$, leading to a contradiction. We now show the converse inclusion $E \subseteq \mathrm{Ext}(B_{\mathcal{M}}(1))$. Let $\mu = y \cdot \delta_\theta$ for some $y \in \mathrm{Ext}(B_{\mathcal{Y}}(1))$ and $\theta \in \Theta$. Suppose there are $t \in (0,1)$ and $\mu_1, \mu_2 \in B_{\mathcal{M}}(1)$ such that (28). We want to show that necessarily $\mu_1 = \mu_2 = \mu$. Consider the subspace

$$\mathcal{Z} = \{z \cdot \delta_\theta : z \in \mathcal{Y}\},$$

and let $\mathcal{P}\colon \mathcal{M}(\Theta, \mathcal{Y}) \to \mathcal{Z}$ be the projection onto $\mathcal{Z}$ defined by

$$\mathcal{P}\nu = \nu(\{\theta\}) \cdot \delta_\theta.$$

By definition of total variation, for every $\nu \in \mathcal{M}(\Theta, \mathcal{Y})$ we have

$$\|\nu\|_{\mathrm{TV}} \geq \|\mathcal{P}\nu\|_{\mathrm{TV}} + \|\nu - \mathcal{P}\nu\|_{\mathrm{TV}},$$

while the converse bound is simply true by triangle inequality, hence

$$\|\nu\|_{\mathrm{TV}} = \|\mathcal{P}\nu\|_{\mathrm{TV}} + \|\nu - \mathcal{P}\nu\|_{\mathrm{TV}}. \tag{29}$$

Note that $\mu \in \mathcal{Z}$ and thus $\mathcal{P}\mu = \mu$. Hence, applying $\mathcal{P}$ to (28) we obtain

$$\mu = t\mathcal{P}\mu_1 + (1-t)\mathcal{P}\mu_2. \tag{30}$$

Now, consider the unit ball in $\mathcal{Z}$,

$$B_{\mathcal{Z}}(1) = \{z \cdot \delta_\theta : z \in B_{\mathcal{Y}}(1)\}.$$

Then $\mu \in \mathrm{Ext}(B_{\mathcal{Z}}(1))$, and $\mathcal{P}\mu_i \in B_{\mathcal{Z}}$ for $i = 1, 2$. Therefore, looking back to (30) we must have $\mathcal{P}\mu_i = \mu$, and in particular $\|\mathcal{P}\mu_i\|_{\mathrm{TV}} = \|\mu\|_{\mathrm{TV}} = 1$. Moreover, $\|\mu_i\|_{\mathrm{TV}} \leq 1$ since $\mu_i \in B_{\mathcal{M}}(1)$. Thus, applying (29) to $\nu = \mu_i$ we get $\|\mu_i - \mathcal{P}\mu_i\|_{\mathrm{TV}} = 0$, hence $\mu_i = \mathcal{P}\mu_i$, which in turn implies $\mu_i = \mu$, concluding the proof. □

The following lemma is contained in the proof of [22, Theorem 3.2]. It allows to reduce a minimization problem over compositional functions to a sequence of interpolation problems with respect to a generic minimizer. Since we found it of independent interest, we thought to emphasize it in a separate lemma.

Let $L$ be a positive integer. Take a set $\mathcal{X}_0$ and Banach spaces $\mathcal{X}_1, \ldots, \mathcal{X}_{L+1}$. For each $\ell = 0, \ldots, L$, fix a Banach space $\mathcal{H}_\ell$ of functions from $\mathcal{X}_\ell$ to $\mathcal{X}_{\ell+1}$. To every $f = f_0 \oplus \cdots \oplus f_L \in \bigoplus_{\ell=0}^{L} \mathcal{H}_\ell$, recall that $f^{\mathrm{deep}} : \mathcal{X}_0 \to \mathcal{X}_{L+1}$ is defined as

$$f^{\mathrm{deep}} = f_L \circ \cdots \circ f_0.$$

**Lemma A.4.** *With the above setting, fix a family $x_1, \ldots, x_N \in \mathcal{X}_0$ and set*

$$\mathcal{A} : \bigoplus_{\ell=0}^{L} \mathcal{H}_\ell \to \mathbb{R}^N, \qquad \mathcal{A}(f)_i = f^{\mathrm{deep}}(x_i).$$

*For $F : \mathbb{R}^N \to (-\infty, +\infty]$, consider the minimization problem*

$$\inf_{f \in \bigoplus_{\ell=0}^{L} \mathcal{H}_\ell} F(\mathcal{A}(f)) + \sum_{\ell=0}^{L} \|f_\ell\|_{\mathcal{H}_\ell}. \tag{31}$$

*Assume that* (31) *has a solution $f^* = \oplus_{\ell=0}^{L} f_\ell^*$, and denote $x_i^{(0)} = x_i$ and $x_i^{(\ell+1)} = f_\ell^*(x_i^{(\ell)})$ for $\ell = 0, \ldots, L$. Then there exists a minimizer $\widetilde{f} = \oplus_{\ell=0}^{L} \widetilde{f}_\ell$ of* (31) *such that for all $\ell = 0, \ldots, L$ the function $\widetilde{f}_\ell$ is the solution of*

$$\inf_{f_\ell \in \mathcal{H}_\ell} \|f_\ell\|_{\mathcal{H}_\ell} \quad \textit{subject to} \quad f_\ell(x_i^{(\ell)}) = x_i^{(\ell+1)}, \qquad i = 1 \ldots, N, \tag{32}$$

*and $\|\widetilde{f}_\ell\|_{\mathcal{H}_\ell} = \|f_\ell^*\|_{\mathcal{H}_\ell}$.*

*Proof.* Let $\widetilde{f}$ be a solution to (32). The solution $f^*$ satisfies the constraints $f_\ell(x_i^{(\ell)}) = x_i^{(\ell+1)}$, hence $\|\widetilde{f}_\ell\|_{\mathcal{H}_\ell} \leq \|f_\ell^*\|_{\mathcal{H}_\ell}$ and $\widetilde{f}_\ell(x_i) = f_\ell^*(x_i)$ for all $i = 1, \ldots, N$. This implies $\sum_{\ell=0}^{L} \|\widetilde{f}_\ell\|_{\mathcal{H}_\ell} \leq \sum_{\ell=0}^{L} \|f_\ell^*\|_{\mathcal{H}_\ell}$ and $\mathcal{A}(\widetilde{f}) = \mathcal{A}(f^*)$, so that

$$F(\mathcal{A}(\widetilde{f})) + \sum_{\ell=0}^{L} \|\widetilde{f}_\ell\|_{\mathcal{H}_\ell} \leq F(\mathcal{A}(f^*)) + \sum_{\ell=0}^{L} \|f_\ell^*\|_{\mathcal{H}_\ell}.$$

But since $f^*$ is a minimizer, so is $\widetilde{f}$ and $\|\widetilde{f}_\ell\|_{\mathcal{H}_\ell} = \|f_\ell^*\|_{\mathcal{H}_\ell}$. □

**Remark A.5.** *We stress that the set $\{x_i^{(\ell)} : \ell = 1, \ldots, L+1, i = 1, \ldots, N\}$ defining the constraints in the problem* (32) *depends on the choice of the minimizer $f^* = \oplus_{\ell=0}^{L} f_\ell$ of the problem* (31)*, so that there is some freedom to choose the points.*

The next lemmas are needed to establish continuity in the setting of our representer theorem 5.3. A continuity result akin to Lemma A.8 is also needed for the proof of [22, Theorem 3.2]. We remark that, in order for the argument to go thorough, *joint* continuity is required, while the proof in [22, Theorem 3.2] only establishes separate continuity. The final result remains valid since joint continuity holds true nevertheless. This can be seen as a special case of our Lemma A.8, where the last steps can be simplified in view of the fact that, in finite-dimensional spaces, weak and strong continuity coincide.

**Lemma A.6.** *Let $\mathcal{X}$ be a set, $\mathcal{Y}$ a Hilbert space, and $\Theta$ a locally compact, second countable topological space. Let $\rho : \mathcal{X} \times \Theta \to \mathbb{R}$ such that $\rho(x, \cdot) \in \mathcal{C}_0(\Theta)$ for all $x \in \mathcal{X}$, and define $\phi(x) : \mathcal{M}(\Theta, \mathcal{Y}) \to \mathcal{Y}$ by*

$$\phi(x)\mu = \int_\Theta \rho(x, \theta) \mathrm{d}\mu(\theta), \qquad x \in \mathcal{X}, \quad \mu \in \mathcal{M}(\Theta, \mathcal{Y}).$$

*Then, for all $x \in \mathcal{X}$, $\phi(x)$ is continuous from $\mathcal{M}(\Theta, \mathcal{Y})$ endowed with the weak$^*$ topology to $\mathcal{Y}$ endowed with the weak topology.*

*Proof.* Note that, since $\mathcal{Y}$ is Hilbert, the Riesz representation theorem says that $\mathcal{M}(\Theta, \mathcal{Y}) = \mathcal{C}_0(\Theta, \mathcal{Y})'$. Now, for all $x \in \mathcal{X}$, $\mu \in \mathcal{M}(\Theta, \mathcal{Y})$ and $y \in \mathcal{Y}$, we have

$$\langle \phi(x)\mu, y \rangle_{\mathcal{Y}} = {}_{\mathcal{C}_0(\Theta,\mathcal{Y})}\langle \rho(x, \cdot)y, \mu \rangle_{\mathcal{M}(\Theta,\mathcal{Y})}.$$

Thus $\langle \phi(x)\cdot, y \rangle_{\mathcal{Y}}$ defines an element of the predual $\mathcal{C}_0(\Theta, \mathcal{Y})$, and hence it is weakly$^*$ continuous from $\mathcal{M}(\Theta, \mathcal{Y})$ to $\mathbb{R}$. But since this is true for all $y \in \mathcal{Y}$, it is weakly$^*$ continuous from $\mathcal{M}(\Theta, \mathcal{Y})$ to $\mathcal{Y}$ endowed with the weak topology. □

The following known result is a direct consequence of Prokhorov theorem. We report the proof for completeness.

**Lemma A.7** (Joint dominated convergence theorem)**.** *Let $\Theta$ be a Polish space. For all $n \in \mathbb{N}$, let $\lambda_n \in \mathcal{M}(\Theta)$ and $f_n : \Theta \to \mathbb{R}$ continuous functions satisfying the following conditions:*

(i) *for each $n \in \mathbb{N}$, the function $f_n$ is $\lambda_n$ integrable;*

(ii) *the sequence $(f_n)_n$ converges to some $f : \Theta \to \mathbb{R}$ uniformly on all compact sets;*

(iii) *the sequence $(f_n)_n$ is uniformly bounded;*

(iv) *the sequence $(\lambda_n)_n$ converges to some $\lambda \in \mathcal{M}(\Theta)$ with respect to the narrow topology,* i.e.

$$\lim_{n\to+\infty} \int_\Theta \varphi(\theta) \mathrm{d}\lambda_n(\theta) = \int_\Theta \varphi(\theta) \mathrm{d}\lambda(\theta), \qquad \forall \varphi \in \mathcal{C}_b(\Theta);$$

(v) *the function $f$ is $\lambda$-integrable.*

*Then*

$$\int_\Theta f_n(\theta) \mathrm{d}\lambda_n(\theta) \xrightarrow[n\to\infty]{} \int_\Theta f(\theta) \mathrm{d}\lambda(\theta).$$

*Proof.* For any compact $K \subset \Theta$, we have

$$\begin{aligned}
&\left| \int_\Theta f_n(\theta) \mathrm{d}\lambda_n(\theta) - \int_\Theta f(\theta) \mathrm{d}\lambda(\theta) \right| \\
&\leq \left| \int_K (f_n(\theta) - f(\theta)) \mathrm{d}\lambda_n(\theta) \right| + \left| \int_{\Theta \setminus K} (f_n(\theta) - f(\theta)) \mathrm{d}\lambda_n(\theta) \right| + \left| \int_\Theta (f_n(\theta) - f(\theta)) \mathrm{d}\lambda(\theta) \right| \\
&+ \left| \int_\Theta f_n(\theta) \mathrm{d}\lambda(\theta) - \int_\Theta f(\theta) \mathrm{d}\lambda_n(\theta) \right| \\
&\leq \sup_K |f_n - f| |\lambda_n|(K) + \sup_{\Theta \setminus K} |f_n - f| |\lambda_n|(\Theta \setminus K) + \left| \int_\Theta (f_n(\theta) - f(\theta)) \mathrm{d}\lambda(\theta) \right| \\
&+ \left| \int_\Theta f_n(\theta) \mathrm{d}\lambda(\theta) - \int_\Theta f(\theta) \mathrm{d}\lambda_n(\theta) \right|.
\end{aligned}$$

The first term goes to zero because $f_n \to f$ uniformly on $K$ and, since $(\lambda_n)_n$ is convergent, $|\lambda_n|(K) \leq |\lambda_n|(\Theta) \leq \sup_n \|\lambda_n\|_{\mathrm{TV}} < \infty$. The third term goes to zero by dominated convergence since $(f_n - f)_n$ is uniformly bounded. The last term goes to zero since both terms converge to the integral $\int_\Theta f(\theta)\mathrm{d}\lambda(\theta)$ as $n \to \infty$: the first term by the dominated convergence theorem since $(f_n)_n$ is uniformly bounded, and the second term by hypothesis $(iv)$ since $f \in C_b(\Theta)$ as uniform limit on compact sets of a uniformly bounded sequence of continuous functions. For the second term, fix an arbitrary $\varepsilon > 0$. Again, $(f_n - f)_n$ is uniformly bounded. Moreover, since $(\lambda_n)_n$ is convergent and $\Theta$ is Polish, by the Prokhorov theorem [6, Theorem 8.6.2] there is a compact set $K_\varepsilon \subset \Theta$ such that $|\lambda_n|(\Theta \setminus K_\epsilon) < \varepsilon$ for all $n$. By taking $K = K_\varepsilon$ we get

$$\limsup_{n\to\infty} \left| \int_\Theta f_n(\theta)\mathrm{d}\lambda_n(\theta) - \int_\Theta f(\theta)\mathrm{d}\lambda(\theta) \right| \leq \left( \sup_n \sup_\Theta |f_n(\theta) - f(\theta)| \right) \epsilon.$$

The claim follows because $\epsilon$ is arbitrary. □

**Lemma A.8.** *Let $\mathcal{X}_0$ be a set, $\mathcal{X}_1, \mathcal{X}_2$ separable Hilbert spaces, and $\Theta_0, \Theta_1$ locally compact, second countable topological spaces. For $\ell = 0, 1$, let $\rho_\ell : \mathcal{X}_\ell \times \Theta_\ell \to \mathbb{R}$ be such that $\rho_\ell(x, \cdot) \in \mathcal{C}_0(\Theta_\ell)$ for all $x \in \mathcal{X}_\ell$, and define $\phi_\ell(x) : \mathcal{M}(\Theta_\ell, \mathcal{X}_{\ell+1}) \to \mathcal{X}_{\ell+1}$ by*

$$\phi_\ell(x)\mu = \int_{\Theta_\ell} \rho_\ell(x,\theta)\mathrm{d}\mu(\theta), \qquad x \in \mathcal{X}_\ell, \qquad \mu \in \mathcal{M}(\Theta_\ell, \mathcal{X}_{\ell+1}).$$

*Assume there are $C > 0$, $g \in \mathcal{C}_b(\Theta_1, \mathcal{X}_1)$ and $\beta \in \mathcal{C}_0(\Theta_1)$ such that, for all $x, x' \in \mathcal{X}_1$ and $\theta \in \Theta_1$,*

$$|\rho_1(x,\theta) - \rho_1(x',\theta)| \leq C|\langle x - x', g(\theta)\rangle_{\mathcal{X}_1}||\beta(\theta)|. \tag{33}$$

*Let $r_0, r_1 > 0$. Then, for all $x \in \mathcal{X}_0$, the map*

$$\Gamma_x(\mu,\nu) = \phi_1((\phi_0(x)\mu))\nu$$

*is jointly weakly* continuous from $B_{\mathcal{M}(\Theta_0,\mathcal{X}_1)}(r_0) \times B_{\mathcal{M}(\Theta_1,\mathcal{X}_2)}(r_1)$ to $\mathcal{X}_2$ endowed with the weak topology.*

*Proof.* By the Banach–Alaoglu theorem, the product $B = B_{\mathcal{M}(\Theta_0,\mathcal{X}_1)}(r_0) \times B_{\mathcal{M}(\Theta_1,\mathcal{X}_2)}(r_1)$ is compact. Moreover, since $\mathcal{X}_{\ell+1}$ ($\ell = 0, 1$) is separable, so is $\mathcal{C}_0(\Theta_\ell, \mathcal{X}_{\ell+1})$ by the Stone–Weierstrass theorem. Also, $\mathcal{X}_{\ell+1}$ is Hilbert, hence $\mathcal{M}(\Theta_\ell, \mathcal{X}_{\ell+1}) = \mathcal{C}_0(\Theta_\ell, \mathcal{X}_{\ell+1})'$ by the Riesz representation theorem. Therefore, $B$ is metrizable [24, Theorem 3.16]. Thus, it is enough to prove the (weak*-weak) *sequential* continuity of $\Gamma_x$.

To this end, let $(\mu_n, \nu_n) \to (\mu, \nu)$ (weakly*). We want to show that $\Gamma_x(\mu_n, \nu_n) \to \Gamma_x(\mu, \nu)$ (weakly). We have

$$\Gamma_x(\mu_n,\nu_n) - \Gamma_x(\mu,\nu) = [\phi_1(\phi_0(x)\mu_n) - \phi_1(\phi_0(x)\mu)]\nu_n + \phi_1(\phi_0(x)\mu)(\nu_n - \nu)\ .$$

The second term goes to zero by Lemma A.6. Let us call $\mathcal{I}$ the first term, and let $z_n = \phi_0(x)\mu_n$, $z = \phi_0(x)\mu$. For all $y \in \mathcal{X}_2$, by assumption (33) we have

$$\begin{aligned} |\langle \mathcal{I}, y\rangle_{\mathcal{X}_2}| &\leq \int_{\Theta_1} |\rho_1(z_n,\theta) - \rho_1(z,\theta)||\beta(\theta)|\mathrm{d}|[\nu_n]_y|(\theta) \\ &\leq C\int_{\Theta_1} |\langle z_n - z, g(\theta)\rangle_{\mathcal{X}_1}||\beta(\theta)|\mathrm{d}|[\nu_n]_y|(\theta), \end{aligned}$$

where $[\nu_n]_y(E) = \langle \nu_n(E), y\rangle_{\mathcal{X}_2}$ for all $E \in \mathcal{B}(\Theta_1)$. We want to apply Lemma A.7 with $f_n(\theta) = |\langle z_n - z, g(\theta)\rangle_{\mathcal{X}_1}|$ and $\lambda_n = |\beta||[\nu_n]_y|$, so to conclude that $|\langle \mathcal{I}, y\rangle_{\mathcal{X}_2}| \to 0$.

First, let us verify that $(f_n)$ is uniformly bounded. We have

$$f_n(\theta) \leq \|z_n - z\|_{\mathcal{X}_1}\|g(\theta)\|_{\mathcal{X}_1} \leq \sup_n \|z_n - z\|_{\mathcal{X}_1}\|g\|_\infty,$$

where $\sup_n \|z_n - z\|_{\mathcal{X}_1} < \infty$ because $(z_n - z)$ is convergent, and $\|g\|_\infty < \infty$ by assumption. Next, we show that $(\lambda_n)$ converges pointwise on $\mathcal{C}_b(\Theta_1)$. Let $\lambda = |\beta||[\nu]_y|$ where $[\nu]_y(E) = \langle \nu(E), y\rangle_{\mathcal{X}_2}$ for all $E \in \mathcal{B}(\Theta_1)$. Then, for every $h \in \mathcal{C}_b(\Theta_1)$ we have $h|\beta| \in \mathcal{C}_0(\Theta_1)$, and hence, since $[\nu_n]_y \to [\nu]_y$ pointwise on $\mathcal{C}_0(\Theta_1)$,

$$\int_{\Theta_1} h(\theta)\mathrm{d}\lambda_n(\theta) = \int_{\Theta_1} h(\theta)|\beta(\theta)|\mathrm{d}[\nu_n]_y(\theta) \to \int_{\Theta_1} h(\theta)|\beta(\theta)|\mathrm{d}[\nu]_y(\theta) = \int_{\Theta_1} h(\theta)\mathrm{d}\lambda(\theta).$$

Finally, we show that $f_n \to 0$ uniformly on compact sets. Let $K \subset \Theta_1$ be compact, and fix an arbitrary $\varepsilon > 0$. Since $g$ is continuous, $g(K)$ is compact in $\mathcal{X}_1$, and thus it can be covered by a finite number of closed balls of radius $\varepsilon$. Let $w_1, \ldots, w_q \in \mathcal{X}_1$ be the centers of such balls, and define $P : \mathcal{X}_1 \to \mathcal{X}_1$ as the projection onto $\operatorname{span}\{w_1, \ldots, w_q\}$. Then $\sup_{w \in g(K)} \|w - Pw\|_{\mathcal{X}_1} \leq \varepsilon$. Hence, for every $\theta \in K$ there is $w \in P\mathcal{X}_1$ such that $\|g(\theta) - w\|_{\mathcal{X}_1} \leq \varepsilon$. Thus, we have

$$\begin{aligned}
|f_n(\theta)| &= |\langle z_n - z, g(\theta)\rangle_{\mathcal{X}_1}| \\
&\leq |\langle z_n - z, g(\theta) - w\rangle_{\mathcal{X}_1}| + |\langle P(z_n - z), w\rangle_{\mathcal{X}_1}| \\
&\leq \|z_n - z\|_{\mathcal{X}_1}\|g(\theta) - w\|_{\mathcal{X}_1} + \|P(z_n - z)\|_{\mathcal{X}_1}\|w\|_{\mathcal{X}_1} \\
&\leq \|z_n - z\|_{\mathcal{X}_1}\|g(\theta) - w\|_{\mathcal{X}_1} + \|P(z_n - z)\|_{\mathcal{X}_1}\left(\|g(\theta) - w\|_{\mathcal{X}_1} + \|g(\theta)\|_{\mathcal{X}_1}\right) \\
&\leq \|z_n - z\|_{\mathcal{X}_1}\varepsilon + \|P(z_n - z)\|_{\mathcal{X}_1}\left(\varepsilon + \|g\|_\infty\right).
\end{aligned}$$

Now, since $z_n \to z$ weakly, $\sup_n \|z_n - z\|_{\mathcal{X}_1} < \infty$, and $\|P(z_n - z)\|_{\mathcal{X}_1} \to 0$ because $P$ has finite rank.

All the assumptions of Lemma A.7 are therefore satisfied, and its application concludes the proof. □

**Remark A.9.** *In the framework of deep integral RKBS, the above lemma states that the evaluation functional at a fixed $x \in \mathcal{X}_0$*

$$(\mu, \nu) \mapsto f_\nu(f_\mu(x))$$

*is jointly weakly* continuous from $B_{\mathcal{M}(\Theta_0, \mathcal{X}_1)}(r_0) \times B_{\mathcal{M}(\Theta_1, \mathcal{X}_2)}(r_1)$ to $\mathcal{X}_2$ endowed with the weak topology. The proof for $L = 2$ can be easily extended to cover the case $L > 2$.*

(F. Bartolucci) ANALYSIS GROUP - DELFT INSTITUTE OF APPLIED MATHEMATICS, TU DELFT, NETHERLANDS
*Email address*: f.bartolucci@tudelft.nl

(E. De Vito) MALGA - DIMA, UNIVERSITY OF GENOA, ITALY
*Email address*: ernesto.devito@unige.it

(L. Rosasco) MALGA - DIBRIS, UNIVERSITY OF GENOA, ITALY & CBMM, MIT & IIT
*Email address*: lorenzo.rosasco@unige.it

(S. Vigogna) ROMADS - DEPARTMENT OF MATHEMATICS, UNIVERSITY OF ROME TOR VERGATA, ITALY
*Email address*: vigogna@mat.uniroma2.it