Harmonic Decompositions of Convolutional Networks

Meyer Scetbon [1 2] **Zaid Harchaoui** [2]

Abstract

We present a description of the function space and the smoothness class associated with a convolutional network using the machinery of reproducing kernel Hilbert spaces. We show that the mapping associated with a convolutional network expands into a sum involving elementary functions akin to spherical harmonics. This functional decomposition can be related to the functional ANOVA decomposition in nonparametric statistics. Building off our functional characterization of convolutional networks, we obtain statistical bounds highlighting an interesting trade-off between the approximation error and the estimation error.

1. Introduction

The renewed interest in convolutional neural networks (Fukushima, 1980; LeCun et al., 1995) in computer vision and signal processing has led to a major leap in generalization performance on common task benchmarks, supported by the recent advances in graphical processing hardware and the collection of huge labelled datasets for training and evaluation. Convolutional neural networks pose major challenges to statistical learning theory. First and foremost, a convolutional network learns from data, jointly, both a feature representation through its hidden layers and a prediction function through its ultimate layer. A convolutional neural network implements a function unfolding as a composition of basic functions (respectively nonlinearity, convolution, and pooling), which appears to model well visual information in images. Yet the relevant function spaces to analyze their statistical performance remain unclear.

The analysis of convolutional neural networks (CNNs) has been an active research topic. Different viewpoints have been developed. A straightforward viewpoint is to dismiss completely the grid- or lattice-structure of images and analyze a multi-layer perceptron (MLP) instead acting on vectorized images, which has the downside to set aside the most interesting property of CNNs which is to model well images that is data with a 2D lattice structure.

The scattering transform viewpoint and the i-theory viewpoint (Mallat, 2012; Bruna & Mallat, 2013; Mallat, 2016; Poggio & Anselmi, 2016; Oyallon et al., 2018) keeps the triad of components nonlinearity-convolution-pooling and their combination in a deep architecture and characterize the group-invariance properties and compression properties of convolutional neural networks. Recent work (Bietti & Mairal, 2017) considers risk bounds involving appropriately defined spectral norms for convolutional kernel networks acting on continuous-domain images.

We present in this paper the construction of a function space including the mapping associated with a convolutional network acting on discrete-domain images. Doing so, we characterize the sequence of eigenvalues and eigenfunctions of the related integral operator, hence shedding light on the harmonic structure of the function space of a convolutional neural network. Indeed the eigenvalue decay controls the statistical convergence rate. Thanks to this spectral characterization, we establish high-probability statistical bounds, relating the eigenvalue decay and the convergence rate.

We show that a convolutional network function admits a decomposition whose structure is related to a functional tensor-product space ANOVA model decomposition (Lin, 2000). Such models extend the popular additive models in order to capture interactions of any order between covariates. Indeed a tensor-product space ANOVA model decomposes a high-dimensional multivariate function as a sum of one-dimensional functions (main effects), two-dimensional functions (two-way interactions), and so on.

A remarkable property of such models is their statistical convergence rate, which is within a log factor of the rate in one dimension, under appropriate assumptions. We bring to light a similar structure in the decomposition of mapping associated with a convolutional network. This structure plays an essential role in the convergence rates we present. This suggests that an important component of the modeling power of a convolutional network is to capture spatial interactions between sub-images or patches.

[1] CREST, ENSAE [2] Department of Statistics, University of Washington. Correspondence to: Meyer Scetbon <meyer.scetbon@ensae.fr>, Zaid Harchaoui <zaid@uw.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

This work makes the following contributions. We construct a kernel and a corresponding reproducing kernel Hilbert space (RKHS) to describe a convolutional network (CNN). The construction encompasses networks with any number of filters per layer. Moreover, we provide a sufficient condition for the kernel to be universal. Then, we establish an explicit, analytical, Mercer decomposition of the multilayer kernel associated to this RKHS. We uncover a relationship to a functional ANOVA model, by highlighting a sum-product structure involving interactions between subimages or patches. We obtain a tight control of the eigenvalue decay of the integral operator associated under general conditions on the activation functions. Finally, we establish convergence rates to the Bayes classifier for the regularized least-squares estimator in this RKHS. From a nonparametric learning viewpoint, these rates are optimal in a minimax sense. All the proofs can be found in the longer version of the paper (Scetbon & Harchaoui, 2020).

2. Basic Notions

Image Space. We first describe the mathematical framework to describe image data. An image is viewed here as a collection of normalized sub-images or patches. The sub-image or patch representation is standard in image processing and computer vision, and encompasses the pixel representation as a special case (Mairal et al., 2014a). Note that the framework presented here readily applies to signals and any grid or lattice-structured data with obvious changes in indexing structures. We focus on the case of images as it is currently a popular application of convolutional networks (Goodfellow et al., 2016).

Denote $\mathcal{X}$ the space of images. Let $h, w \geq 1$ respectively the height and width of the images and $\min(h^2, w^2) \geq d \geq 2$ the size of each patch. We consider square patches for simplicity. Denoting $r \geq 1$ the height of a patch, we have that $r^2 = d$. We define for each $(i, j) \in \{1, ..., h-r+1\} \times \{1, ..., w-r+1\}$ the patch extraction operator at location (i, j) as

$$P_{i,j}(\mathbf{X}) := (\mathbf{X}_{i+\ell, j+k})_{\ell, k \in \{1,...,r\}} \in \mathbb{R}^d \qquad (1)$$

where $\mathbf{X} \in \mathbb{R}^{h \times w}$. Moreover let $1 \leq n \leq (h-r+1)(w-r+1)$ and let $A \subset \{1, ..., h-r+1\} \times \{1, ..., w-r+1\}$ such that $|A| = n$.

Define now the initial space of images as $E_A := \{\mathbf{X} \in \mathbb{R}^{h \times w} \colon \ \|P_z(\mathbf{X})\|_2 = 1 \ \text{ for } \ z \in A\}$ where each patch considered has been normalized. Since $\{(i+\ell, j+k) \colon \ (i,j) \in A \ \text{ and } \ \ell, k \in \{1, ..., r\}\} = \{1, ..., h-r+1\} \times \{1, ..., w-r+1\}$, the mapping

$$\begin{array}{cccc} \phi & : & \mathbb{R}^{h \times w} & \rightarrow & \mathbb{R}^d \times ... \times \mathbb{R}^d \\ & & \mathbf{X} & \rightarrow & (P_z(\mathbf{X}))_{z \in A} \end{array}$$

is injective. The mappings $\mathcal{I} := \phi(E_A)$ and E_A are then isomorphic. Hence we shall work from now on with $\mathcal{I}$ as the image space.

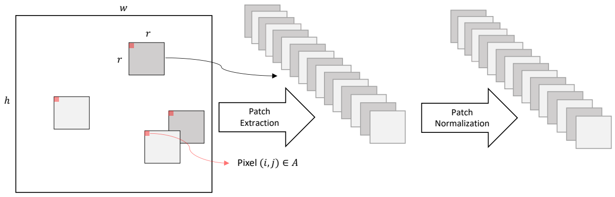

Figure 1. For simplicity, we consider a single-channel image in this illustration. Normalized patches are extracted from the image.

We have by construction that $\mathcal{I} \subset \prod_{i=1}^{n} S^{d-1}$ the n-th Cartesian power of S^{d-1}, where S^{d-1} is the unit sphere of $\mathbb{R}^d$. Moreover, as soon as the patches considered are disjoint, we have that $\mathcal{I} = \prod_{i=1}^{n} S^{d-1}$. In order to simplify the notation, we shall always consider the case where $\mathcal{I} = \prod_{i=1}^{n} S^{d-1}$ where d is the dimension of the square patches and n is the number of patches considered. In the following, we shall denote for any $q \geq 1$ and set $\mathcal{X}$, the q-ary Cartesian power $\prod_{i=1}^{q} \mathcal{X} := (\mathcal{X})^q$. Moreover if $\mathbf{X} \in (\mathcal{X})^q$, we shall denote $\mathbf{X} := (\mathbf{X}_i)_{i=1}^{q}$ where each $\mathbf{X}_i \in \mathcal{X}$.

Let $P_m(d)$ be the space of homogeneous polynomials of degree m in d variables with real coefficients and $\mathcal{H}_m(d)$ be the space of harmonics polynomials defined by

$$\mathcal{H}_m(d) := \{P \in P_m(d) | \Delta P = 0\} \qquad (2)$$

where $\Delta \cdot = \sum_{i=1}^{d} \frac{\partial^2 \cdot}{\partial x_i^2}$ is the Laplace operator on $\mathbb{R}^d$ (Folland, 2016).

Moreover, define $H_m(S^{d-1})$ the space of real spherical harmonics of degree m defined as the set of restrictions of harmonic polynomials in $\mathcal{H}_m(d)$ to S^{d-1}. Let also $L_2^{d\sigma_{d-1}}(S^{d-1})$ be the space of (real) square-integrable functions on the sphere S^{d-1} endowed with its induced Lebesgue measure $d\sigma_{d-1}$ and $|S^{d-1}|$ the surface area of S^{d-1}. $L_2^{d\sigma_{d-1}}(S^{d-1})$ endowed with its natural inner product is a separable Hilbert space and the family of spaces $(H_m(S^{d-1}))_{m \geq 0}$, yields a direct sum decomposition (Efthimiou & Frye, 2014)

$$L_2^{d\sigma_{d-1}}(S^{d-1}) = \bigoplus_{m \geq 0} H_m(S^{d-1}) \qquad (3)$$

which means that the summands are closed and pairwise orthogonal. Moreover, each $H_m(S^{d-1})$ has a finite dimension $\alpha_{m,d}$ with $\alpha_{0,d} = 1$, $\alpha_{1,d} = d$ and for $m \geq 2$

$$\alpha_{m,d} = \binom{d-1+m}{m} - \binom{d-1+m-2}{m-2}$$

Therefore for all $m \geq 0$, given any orthonormal basis of $H_m(S^{d-1})$, $(Y_m^1, ..., Y_m^{\alpha_{m,d}})$, we can build an Hilbertian basis of $L_2^{d\sigma_{d-1}}(S^{d-1})$ by concatenating these orthonormal basis. Let $L_2(\mathcal{I}) := L_2^{\otimes_{i=1}^n d\sigma_{d-1}}(\mathcal{I})$ be the space of (real) square-integrable functions on $\mathcal{I}$ endowed with the n-tensor product measure $\otimes_{i=1}^n d\sigma_{d-1} := d\sigma_{d-1} \otimes ... \otimes d\sigma_{d-1}$ and let us define the integral operator on $L_2(\mathcal{I})$ associated with a positive semi-definite kernel K on $\mathcal{I}$

$$\begin{array}{rccc} T_K & : & L_2(\mathcal{I}) & \to & L_2(\mathcal{I}) \\ & & f & \to & \int_{\mathcal{I}} K(x,\cdot) f(x) \otimes_{i=1}^n d\sigma_{d-1}(x). \end{array}$$

As soon as $\int_{\mathcal{I}} K(x,x) d\sigma_{d-1} \otimes ... \otimes d\sigma_{d-1}(x)$ is finite, which is clearly satisfied when K is continuous, T_K is well defined, self-adjoint, positive semi-definite and trace-class (Simon, 2010; Smale & Zhou, 2007).

We approach here the modeling of interactions of patches or sub-images with functional ANOVA modelling in mind. Let us first recall the basic notions to define a tensor product of functional Hilbert spaces (Lin, 2000; Steinwart & Christmann, 2008). Consider a Hilbert space E_1 of functions of $\mathbf{X}_1$ and a Hilbert space E_2 of functions of $\mathbf{X}_2$. The tensor product space $E_1 \otimes E_2$ is defined as the completion of the class of functions of the form

$$\sum_{i=1}^{k} f_i(\mathbf{X}_1) g_i(\mathbf{X}_2)$$

where $f_i \in E_1$, $g_i \in E_2$ and k is any positive integer, under the norm induced by the norms in E_1 and E_2. The inner product in $E_1 \otimes E_2$ satisfies

$$\begin{aligned} &\langle f_1(\mathbf{X}_1) g_1(\mathbf{X}_2), f_2(\mathbf{X}_1) g_2(\mathbf{X}_2) \rangle_{E_1 \otimes E_2} \\ &\quad = \langle f_1(\mathbf{X}_1), f_2(\mathbf{X}_1) \rangle_{E_1} \langle g_1(\mathbf{X}_2), g_2(\mathbf{X}_2) \rangle_{E_2} \end{aligned}$$

where for $i = 1, 2$, $\langle \cdot, \cdot \rangle_{E_i}$ denote the inner product in E_i. Note that when E_1 and E_2 are RKHS-s with associated kernels k_1 and k_2, one has an explicit formulation of the kernel associated to the RKHS $E_1 \otimes E_2$ (Carmeli et al., 2010). A tensor product space ANOVA model captures interactions between covariates as follows. Let D be the highest order of interaction in the model. Such model assumes that the high-dimensional function to be estimated is a sum of one-dimensional functions (main effects), two-dimensional functions (two-way interactions), and so on. That is, the n-dimensional function f decomposes as

$$f(x_1, ..., x_n) = C + \sum_{k=1}^{D} \sum_{\substack{A \subset \{1,...,n\} \\ |A| = k}} f_A(x_A)$$

where C is a constant and the components satisfy conditions that guarantee the uniqueness (Scetbon & Harchaoui, 2020). More precisely, after assigning a function space for each main effect, this strategy models an interaction as living in the tensor product space of the function spaces of the interacting main effects. In other words, if we assume $f_1(\mathbf{X}_1)$ to be in a Hilbert space E_1 of functions of $\mathbf{X}_1$ and $f_2(\mathbf{X}_2)$ be in a Hilbert space E_2 of functions of $\mathbf{X}_2$, then we can model f_{12} as in $E_1 \otimes E_2$, the tensor product space of E_1 and E_2. Higher order interactions are modeled similarly. In (Lin, 2000), the author considers the case where the main effects are univariate functions living in a Sobolev–Hilbert space with order $m \geq 1$ and domain $[0, 1]$, denoted $H^m([0, 1])$, defined as

$$\left\{ f\colon f^{(\nu)} \text{ abs. cont.}, \nu = 0, ..., m-1; f^{(m)} \in L^2 \right\}$$

More generally, functional ANOVA models assume that the main effects are univariate functions living in a RKHS (Lin, 2000).

3. Convolutional Networks and Multi-Layer Kernels

We proceed with the mathematical description of a convolutional network. The description follows previous works (Bruna & Mallat, 2013; Mairal et al., 2014b; Bietti & Mairal, 2017). Let N be the number of hidden layers, $(\sigma_i)_{i=1}^N$, N real-valued functions defined on $\mathbb{R}$ be the activation functions at each hidden layer, $(d_i)_{i=1}^N$ the sizes of square patches at each hidden layer, $(p_i)_{i=1}^N$ the number of filters at each hidden layer and $(n_i)_{i=1}^{N+1}$ the number of patches at each hidden layer. As our input space is $\mathcal{I} = (S^{d-1})^n$, we set $d_0 = d$, $p_0 = 1$, $n_0 = n$. Moreover as the prediction layer is a linear transformation of the N^{th} layer, we do not need to extract patches from the N^{th} layer, and we set $d_N = n_{N-1}$ such that the only "patch" extracted for the prediction layer is the full "image" itself. Therefore we can also set $n_N = 1$.

Then, a mapping defined by a convolutional network is parameterized by a sequence $W := (W^0, ..., W^N)$ where for $0 \leq k \leq N-1$, $W^k \in \mathbb{R}^{p_{k+1} \times d_k p_k}$ and $W^N \in \mathbb{R}^{d_N p_N}$ for the prediction layer. Indeed let W such a sequence and denote for $k \in \{0, ..., N-1\}$, $W^k := (w_1^k, ..., w_{p_{k+1}}^k)^T$ where for all $j \in \{1, ..., p_{k+1}\}$, $w_j^k \in \mathbb{R}^{d_k p_k}$. Moreover let us define for all $k \in \{0, ..., N-1\}$, $j \in \{1, ..., p_{k+1}\}$ and $q \in \{1, ..., n_{k+1}\}$ the following sequence of operators.

Convolution Operators.

$$C_j^k : \mathbf{Z} \in (\mathbb{R}^{d_k p_k})^{n_k} \longrightarrow C_j^k(\mathbf{Z}) := \left(\langle \mathbf{Z}_i, w_j^k \rangle \right)_{i=1}^{n_k} \in \mathbb{R}^{n_k}$$

Non-Linear Operators.

$$M_k : \mathbf{X} \in \mathbb{R}^{n_k} \longrightarrow M_k(\mathbf{X}) := \left(\sigma_k\left(\mathbf{X}_i\right) \right)_{i=1}^{n_k} \in \mathbb{R}^{n_k}$$

Pooling Operators. Let $(\gamma_{i,j}^k)_{i,j=1}^{n_k}$ be the pooling factors at layer k (which are often assumed to be decreasing with respect to the distance between i and j).

$$A_k : \mathbf{X} \in \mathbb{R}^{n_k} \longrightarrow A_k(\mathbf{X}) := \left(\sum_{j=1}^{n_k} \gamma_{i,j}^k \mathbf{X}_j\right)_{i=1}^{n_k} \in \mathbb{R}^{n_k}$$

Patch extraction Operators.

$$\begin{array}{rccc} P_q^{k+1} & : \quad (\mathbb{R}^{p_{k+1}})^{n_k} & \to & \mathbb{R}^{p_{k+1} d_{k+1}} \\ & \mathbf{U} & \to & P_q^{k+1}(\mathbf{U}) := (\mathbf{U}_{q+l})_{\ell=0}^{d_{k+1}-1} \end{array}$$

Notice that, as we set $d_N = n_{N-1}$ and $n_N = 1$, hence when $k = N-1$, there is only one patch extraction operator which is $P_1^N = \text{Id}$.

Then the function associated to W generated by the convolutional network can be obtained by the following procedure: let $\mathbf{X}^0 \in \mathcal{I}$, then we can denote $\mathbf{X}^0 = (\mathbf{X}_i^1)_{i=1}^{n_1}$ where for all $i \in [|1, n_1|]$, $\mathbf{X}_i^0 \in S^{d-1}$. Therefore we can build by induction the sequence $(\mathbf{X}^k)_{k=0}^N$ by doing the following operations starting from $k = 0$ until $k = N-1$

$$C_j^k(\mathbf{X}^k) = \left(\langle \mathbf{X}_i^k, w_j^k \rangle\right)_{i=1}^{n_k} \tag{4}$$

$$M_k(C_j^k(\mathbf{X}^k)) = \left(\sigma_k\left(\langle \mathbf{X}_i^k, w_j^k \rangle\right)\right)_{i=1}^{n_k} \tag{5}$$

$$A_k(M_k(C_j^k(\mathbf{X}^k))) = \left(\sum_{q=1}^{n_k} \gamma_{i,q}^k \sigma_k\left(\langle \mathbf{X}_q^k, w_j^k \rangle\right)\right)_{i=1}^{n_k} \tag{6}$$

Writing now

$$\mathbf{Z}^{k+1}(i,j) = A_k(M_k(C_j^k(\mathbf{X}^k)))_i$$

we have

$$\hat{\mathbf{X}}^{k+1} = (\mathbf{Z}_{k+1}(i,1), ..., \mathbf{Z}_{k+1}(i, p_{k+1})))_{i=1}^{n_k}$$

and finally

$$\mathbf{X}^{k+1} = (P_q^{k+1}(\hat{\mathbf{X}}^{k+1}))_{q=1}^{n_{k+1}} \tag{7}$$

The mapping associated with a convolutional network therefore reads $\mathcal{N}_W(\mathbf{X}^0) := \langle \mathbf{X}^N, W^N \rangle_{\mathbb{R}^{p_N n_{N-1}}}$. In the following, we denote $\mathcal{F}_{(\sigma_i)_{i=1}^N, (p_i)_{i=1}^N}$ the function space of all the functions $\mathcal{N}_W$ defined as above on $\mathcal{I}$ for any choice of $(W^k)_{k=0}^N$ such that for $0 \leq k \leq N-1$, $W^k \in \mathbb{R}^{p_{k+1} \times d_k p_k}$ and $W^N \in \mathbb{R}^{d_N\, p_N}$. We omit the dependence of $\mathcal{F}_{(\sigma_i)_{i=1}^N, (p_i)_{i=1}^N}$ with respect to $(d_i)_{i=1}^N$ and $(n_i)_{i=1}^N$ to simplify the notations. We shall also consider the union space

$$\mathcal{F}_{(\sigma_i)_{i=1}^N} := \bigcup_{(p_1, ..., p_N) \in \mathbb{N}_*{}^N} \mathcal{F}_{(\sigma_i)_{i=1}^N, (p_i)_{i=1}^N}$$

to encompass convolutional networks with varying number of filters across layers.

Example. Consider the case where at each layer the number of filters is 1. This corresponds to the case where for all $k \in \{1, ..., N\}$, $p_k = 1$. Therefore we can omit the dependence in j of the convolution operators defined above. At each layer k, $\widehat{\mathbf{X}}^{k+1} \in \mathbb{R}^{n_k}$ is the new image obtained after a convolution, a nonlinear and a pooling operation with n_k pixels which is the number of patches that has been extracted from the image $\widehat{\mathbf{X}}^k$ at layer $k-1$. Moreover $\mathbf{X}^{k+1}$ is the decomposition of the image $\hat{\mathbf{X}}^{k+1}$ in n_{k+1} patches obtained thanks to the patch extraction operators $(P_q^{k+1})_{q=1}^{n_{k+1}}$.

Finally after N layers, we obtain that $\hat{\mathbf{X}}^N = \mathbf{X}^N \in \mathbb{R}^{n_{N-1}}$ which is the final image with n_N pixels obtained after repeating N times the above operations. Then the prediction layer is a linear combination of the coordinates of the final image $\mathbf{X}^N$ from which we can finally define for all $\mathbf{X}^0 \in \mathcal{I}$, $\mathcal{N}_W(\mathbf{X}^0) := \langle \mathbf{X}^N, W^N \rangle_{\mathbb{R}^{n_{N-1}}}$.

We show in Prop. 1 below that there exists an RKHS (Schölkopf & Smola, 2002) containing the space of functions $\mathcal{F}_{(\sigma_i)_{i=1}^N}$, and this, for a general class of activation functions, $(\sigma_i)_{i=1}^N$, admitting a Taylor decomposition on $\mathbb{R}$. Moreover we show that for a large class of nonlinear functions, the kernel is actually a c-universal kernel on $\mathcal{I}$. It is worthwhile to emphasize that the definition of the RKHS H_N we give below does not depend on the number of filters $(p_i)_{i=2}^{N+1}$ at each hidden layer. Therefore our framework encompasses networks with varying number of filters across layers (Scetbon & Harchaoui, 2020).

Definition 3.1. *(c-**universal** (Sriperumbudur et al., 2011)) A continuous positive semi-definite kernel k on a compact Hausdorff space $\mathcal{X}$ is called c-universal if the RKHS, H induced by k is dense in $\mathcal{C}(\mathcal{X})$ w.r.t. the uniform norm.*

Proposition 1. *Let $N \geq 2$ and $(\sigma_i)_{i=1}^N$ be a sequence of N functions with a Taylor decomposition on $\mathbb{R}$. Moreover let $(f_i)_{i=1}^N$ be the sequence of functions such that for every $i \in \{1, ..., N\}$*

$$f_i(x) = \sum_{t \geq 0} \frac{|\sigma_i^{(t)}(0)|}{t!} x^t \tag{8}$$

Then the bivariate function defined on $\mathcal{I} \times \mathcal{I}$ as

$$K_N(\mathbf{X}, \mathbf{X}') := f_N \circ ... \circ f_2 \left(\sum_{i=1}^{n} f_1\left(\langle \mathbf{X}_i, \mathbf{X}'_i \rangle_{\mathbb{R}^d}\right)\right)$$

is a positive definite kernel on $\mathcal{I}$, and the RKHS associated H_N contains $\mathcal{F}_{(\sigma_i)_{i=1}^N}$, the function space generated by convolutional networks. Moreover as soon as $\sigma_i^{(t)}(0) \neq 0$ for all $i \geq 1$ and $t \geq 0$, then K_N is a c-universal kernel on $\mathcal{I}$.

Function space. A simple fact is that

$$\inf_{f \in H_N} \mathbb{E}[(f(X) - Y)^2] \leq \inf_{f \in F} \mathbb{E}[(f(X) - Y)^2]$$

where $F := \mathcal{F}_{(\sigma_i)_{i=1}^N}$. In other words the minimum expected risk in H_N is a lower bound on the minimum expected risk in F. Since a major concern of recent years has been the spectacular performance of deep networks *i.e.* how small they can drive the risk, analyzing them via this kind of Hilbertian envelope can shed more light on the relation between their multi-layer structure and their statistical performance. From this simple observation, one could obtain statistical bounds on F using high-probability bounds from (Boucheron et al., 2005). However, we choose to focus on getting tight statistical bounds on H_N instead, in order to explore the connection between the statistical behavior and the integral operator eigenspectrum.

Universality. For a large class of nonlinear activation functions, the kernel K_N defined above is actually *universal*. Therefore the corresponding RKHS H_N allows one to get universal consistency results for common loss functions (Steinwart & Christmann, 2008). In particular, if we choose the least-squares loss, we have then

$$\inf_{f\in H} \mathbb{E}[(f(X)-Y)^2] = R^\star$$

where $R^\star$ is the Bayes risk. See Corollary 5.29 in (Steinwart & Christmann, 2008).

For instance, if at each layer the nonlinear function is $\sigma_{\text{exp}}(x) = \exp x$, as in (Mairal et al., 2014b; Bietti & Mairal, 2017), then the corresponding RKHS is universal. There are other examples of activation functions satisfying assumptions from Prop. 1, such as the square activation $\sigma_2(x) = x^2$, the smooth hinge activation σ_{sh}, close to the ReLU activation, or a sigmoid-like function such as σ_{erf}, similar to the sigmoid function, with

$$\sigma_{\text{erf}}(x) = \frac{1}{2}\left(1 + \frac{1}{\sqrt{\pi}}\int_{-\sqrt{\pi}x}^{\sqrt{\pi}x} e^{-t^2}\,dt\right)$$
$$\sigma_{\text{sh}}(x) = \frac{1}{\sqrt{\pi}}\int_{-x}^{x} xe^{-t^2}\,dt + \frac{\exp(-\pi x^2)}{2\pi}$$

In the following section, we study in detail the properties of the kernel K_N. In particular we show an explicit Mercer decomposition of the kernel from which we uncover a relationship existing between convolutional networks and functional ANOVA models.

4. Spectral Analysis of Convolutional Networks

We give now a Mercer decomposition of the kernel introduced in Prop 1. From this Mercer decomposition, we show first that the multivariate function space generated by a convolutional network enjoys a decomposition related to the one in functional ANOVA models, where the highest order of interaction is controlled by the nonlinear functions $(\sigma_i)_{i=1}^N$ involved in the construction of the network. Moreover we also obtain a tight control of the eigenvalue decay under general assumptions on the activation functions involved in the construction of the network.

Recall that for all $m \geq 0$, we denote $(Y_m^1, ..., Y_m^{\alpha_{m,d}})$ an arbitrary orthonormal basis of $H_m(S^{d-1})$. The next result gives an explicit Mercer decomposition of the kernels of interest.

Theorem 4.1. *Let $N \geq 2$, f_1 a real valued function that admits a Taylor decomposition around 0 on $[-1, 1]$ with non-negative coefficients and $(f_i)_{i=2}^N$ a sequence of real valued functions such that $f_N \circ ... \circ f_2$ admits a Taylor decomposition around 0 on $\mathbb{R}$ with non-negative coefficients $(a_q)_{q\geq 0}$. Let us denote for all $k_1, ..., k_n \geq 0$, $(l_{k_1}, ..., l_{k_n}) \in \{1, ..., \alpha_{k_1,d}\} \times ... \times \{1, ..., \alpha_{k_n,d}\}$ and $\mathbf{X} \in \mathcal{I}$,*

$$e_{(k_i,l_{k_i})_{i=1}^n}(\mathbf{X}) := \prod_{i=1}^n Y_{k_i}^{l_{k_i}}(\mathbf{X}_i)$$

Then each $e_{(k_i,l_{k_i})_{i=1}^n}$ is an eigenfunction of T_{K_N} the integral operator associated to the kernel K_N, with associated eigenvalue given by the formula

$$\mu_{(k_i,l_{k_i})_{i=1}^n} := \sum_{q\geq 0} a_q \sum_{\substack{\alpha_1,...,\alpha_n \geq 0 \\ \sum_{i=1}^n \alpha_i = q}} \binom{q}{\alpha_1, ..., \alpha_n} \prod_{i=1}^n \lambda_{k_i,\alpha_i}$$

where for any $k \geq 0$ and $\alpha \geq 0$ we have

$$\lambda_{k,\alpha} = \frac{|S^{d-2}|\Gamma((d-1)/2)}{2^{k+1}}$$
$$\sum_{s\geq 0}\left[\frac{d^{2s+k}}{dt^{2s+k}}|_{t=0}\frac{f_1^\alpha(t)}{(2s+k)!}\right]\frac{(2s+k)!}{(2s)!}\frac{\Gamma(s+1/2)}{\Gamma(s+k+d/2)}$$

Moreover we have

$$K_N(\mathbf{X}, \mathbf{X}') =$$
$$\sum_{\substack{k_1,...,k_n\geq 0 \\ 1\leq l_{k_i}\leq \alpha_{k_i,d}}} \mu_{(k_i,l_{k_i})_{i=1}^n} e_{(k_i,l_{k_i})_{i=1}^n}(\mathbf{X}) e_{(k_i,l_{k_i})_{i=1}^n}(\mathbf{X}')$$

where the convergence is absolute and uniform.

From this Mercer decomposition, we get a decomposition of the multivariate function generated by a convolutional network. This decomposition is related to the one in functional ANOVA models. But first let us introduce useful notations. Let us denote

$$L_2^{d\sigma_{d-1}}(S^{d-1}) = \{1\} \bigoplus L_{2,0}^{d\sigma_{d-1}}(S^{d-1})$$

where $L_{2,0}^{d\sigma_{d-1}}(S^{d-1})$ is the subspace orthogonal to $\{1\}$. Thus we have

$$\bigotimes_{i=1}^n L_2^{d\sigma_{d-1}}(S^{d-1}) = \bigotimes_{i=1}^n [\{1\} \bigoplus L_{2,0}^{d\sigma_{d-1}}(S^{d-1})].$$

Identify the tensor product of $\{1\}$ with any Hilbert space with that Hilbert space itself, then $\bigotimes_{i=1}^{n} L_2^{d\sigma_{d-1}}(S^{d-1})$ is the direct sum of all the subspaces of the form $L_{2,0}^{d\sigma_{d-1}}(\mathbf{X}_{j_1}) \otimes ... \otimes L_{2,0}^{d\sigma_{d-1}}(\mathbf{X}_{j_k})$ and $\{1\}$ where $\{j_1, ..., j_k\}$ is a subset of $\{1, ..., n\}$ and the subspaces in the decomposition are all orthogonal to each other.

In fact, the function space generated by a convolutional network is a subset of $\bigotimes_{i=1}^{n} L_2^{d\sigma_{d-1}}(S^{d-1})$ which selects only few orthogonal components in the decomposition of $\bigotimes_{i=1}^{n} L_2^{d\sigma_{d-1}}(S^{d-1})$ described above and allows only few interactions between the covariates. Moreover, the highest order of interactions can be controlled by the depth of the network. Indeed, in the following proposition, we show that the eigenvalues $(\mu_{(k_i, l_{k_i})_{i=1}^n})$ obtained from the Mercer decomposition vanish as soon as the interactions are large enough relatively to the network depth (Scetbon & Harchaoui, 2020).

Proposition 4.1. *Let $N \geq 2$, f_1 a real-valued function admitting a Taylor decomposition around 0 on $[-1, 1]$ with non-negative coefficients and $f_N \circ .... \circ f_2$ a polynomial of degree $D \geq 1$. Then, denoting $d^* := \min(D, n)$, we have that $\mu_{(k_i, l_{k_i})_{i=1}^n} = 0$, as soon as $|\{i : k_i \neq 0\}| > d^*$, and, for any $f \in \mathcal{F}_{(\sigma_i)_{i=1}^n}$, $q > d^*$ and $\{j_1, ..., j_q\} \subset \{1, ..., n\}$, we have*

$$f \in \left(L_{2,0}^{d\sigma_{d-1}}(\mathbf{X}_{j_1}) \otimes ... \otimes L_{2,0}^{d\sigma_{d-1}}(\mathbf{X}_{j_q})\right)^{\perp}.$$

From this observation, we are able to characterize the function space of a convolutional network following the same strategy as functional ANOVA models, but allowing the main effects to live in an Hilbert space which may not necessarily be a RKHS of univariate functions. We shall refer to such decompositions as ANOVA-like decompositions to underscore both their similarity and their difference with functional ANOVA decompositions.

Definition 4.1. ***ANOVA-like Decomposition*** *Let f a real valued function defined on $\mathcal{I}$. We say that f admits an ANOVA-like decomposition of order r if f can be written as*

$$f(\mathbf{X}_1, ..., \mathbf{X}_n) = C + \sum_{k=1}^{r} \sum_{\substack{A \subset \{1,...,n\} \\ |A|=k}} f_A(\mathbf{X}_A)$$

where C is a constant, for all $k \in \{1, r\}$ and for $A = \{j_1, ..., j_k\} \subset \{1, ..., n\}$ $\mathbf{X}_A = (\mathbf{X}_{j_1}, ..., \mathbf{X}_{j_k})$, $f_A \in L_{2,0}^{d\sigma_{d-1}}(\mathbf{X}_{j_1}) \otimes ... \otimes L_{2,0}^{d\sigma_{d-1}}(\mathbf{X}_{j_k})$ and the decomposition is unique.

Here the main effects live in $L_{2,0}^{d\sigma_{d-1}}(S^{d-1})$ which is a Hilbert space of multivariate functions. Thanks to Prop. 4.1, any function generated by a convolutional network admits an ANOVA-like decomposition, where the highest order of interactions is at most d^* and it is completely determined by the functions $(\sigma_i)_{i=1}^N$. Moreover even if the degree D is arbitrarily large, the highest order of interaction cannot be bigger than n.

Example. For any convolutional network such that σ_1 admits a Taylor decomposition around 0 on $[-1, 1]$ (as in $\tanh$ and other nonlinear activations), and $(\sigma_i)_{i=2}^N$, are quadratic functions, the highest order of interaction allowed by the network is upper bounded by $d^* \leq \min(2^{N-1}, n)$.

Finally, from the Mercer decomposition, we can control the eigenvalue decay under general assumptions on the activations. Assume that $N \geq 2$ and that f_1 is a function, admitting a Taylor decomposition on $[-1, 1]$, with non-negative coefficients $(b_m)_{m \geq 0}$. Let us now show a first control of the $(\lambda_{m,\alpha})_{m,\alpha}$ introduced in Theorem 4.1.

Proposition 4.2. *If there exist $1 > r > 0$ and $0 < c_2 \leq c_1$ constants such that for all $m \geq 0$*

$$c_2 r^m \leq b_m \leq c_1 r^m$$

then for all $\alpha \geq 1$, there exist $C_{1,\alpha}$ and $C_{2,\alpha} > 0$, constants depending only on α, and d such that for all $m \geq 0$

$$C_{2,\alpha} \left(\frac{r}{4}\right)^m \leq \lambda_{m,\alpha} \leq C_{1,\alpha}(m+1)^{\alpha-1} r^m$$

We can now provide a tight control of the positive eigenvalues of the integral operator T_{K_N} associated with the kernel K_N sorted in a non-increasing order with their multiplicities which is exactly the ranked sub sequence of positive eigenvalues in $\left(\mu_{(k_i, l_{k_i})_{i=1}^n}\right)$.

Proposition 4.3. *Let us assume that $f_N \circ .... \circ f_2$ is a polynomial of degree $D \geq 1$ and let $d^* := \min(D, n)$. Let $(\mu_m)_{m=0}^M$ be the positive eigenvalues of the integral operator T_{K_N} associated to the kernel K_N ranked in a non-increasing order with their multiplicities, where $M \in \mathbb{N} \cup \{+\infty\}$. Under the same assumptions of Prop. 4.2 we have $M = +\infty$ and there exists $C_3, C_4 > 0$ and $0 < \gamma < q$ constants such that for all $m \geq 0$:*

$$C_4 e^{-q m^{\frac{1}{(d-1)d^*}}} \leq \mu_m \leq C_3 e^{-\gamma m^{\frac{1}{(d-1)d^*}}}$$

Thanks to this control, we obtain in the next section the convergence rate for the regularized least-squares estimator on a non-trivial set of distributions, for the class of RKHS introduced earlier. Moreover, in some situations, we show that these convergence rates are actually minimax optimal from a nonparametric learning viewpoint.

5. Regularized Least-Squares for CNNs

We consider the standard nonparametric learning framework (Györfi et al., 2002; Steinwart & Christmann, 2008), where the goal is to learn, from independent and identically distributed examples $\mathbf{z} = \{(x_1, y_1), \dots, (x_\ell, y_\ell)\}$ drawn from an unknown distribution ρ on $Z := \mathcal{I} \times \mathcal{Y}$, a functional dependency $f_\mathbf{z} : \mathcal{I} \to \mathcal{Y}$ between input $x \in \mathcal{I}$ and output $y \in \mathcal{Y}$. The joint distribution $\rho(x, y)$, the marginal distribution $\rho_\mathcal{I}$, and the conditional distribution $\rho(.|x)$, are related through $\rho(x, y) = \rho_\mathcal{I}(x)\rho(y|x)$. We call $f_\mathbf{z}$ the learning method or the estimator and the learning algorithm is the procedure that, for any sample size $\ell \in \mathbb{N}$ and training set $\mathbf{z} \in Z^\ell$ yields the learned function or estimator $f_\mathbf{z}$. Here we assume that $\mathcal{Y} \subset \mathbb{R}$, and given a function $f : \mathcal{I} \to \mathcal{Y}$, the ability of f to describe the distribution ρ is measured by its expected risk

$$R(f) := \int_{\mathcal{I}\times\mathcal{Y}} (f(x) - y)^2 \, d\rho(x, y) \ . \tag{9}$$

The minimizer over the space of measurable $\mathcal{Y}$-valued functions on $\mathcal{I}$ is

$$f_\rho(x) := \int_\mathcal{Y} y d\rho(y|x) \ . \tag{10}$$

We seek to characterize, with high probability, how close $R(f_\mathbf{z})$ is to $R(f_\rho)$. Let us now consider the regularized least-squares estimator (RLS). Consider as hypothesis space a Hilbert space H of functions $f : \mathcal{I} \to \mathcal{Y}$. For any regularization parameter $\lambda > 0$ and training set $\mathbf{z} \in Z^\ell$, the RLS estimator $f_{H,\mathbf{z},\lambda}$ is the solution of

$$\min_{f \in H} \left\{ \frac{1}{\ell} \sum_{i=1}^{\ell} (f(x_i) - y_i)^2 + \lambda \|f\|_H^2 \right\} \ . \tag{11}$$

In the following we consider the specific estimators obtained from the RKHS-s introduced in Prop. 1. But before stating the statistical bounds we have obtained, we recall basic definitions in order to clarify what we mean by asymptotic upper rate, lower rate and minimax rate optimality, following (Caponnetto & De Vito, 2007). We want to track the precise behavior of these rates and the effects of adding layers in a convolutional network. More precisely, we consider a class of Borel probability distributions $\mathcal{P}$ on $\mathcal{I} \times \mathbb{R}$ satisfying basic general assumptions. We consider rates of convergence according to the $L_2^{d\rho_\mathcal{I}}$ norm denoted $\|.\|_\rho$.

Definition 5.1. *(Upper Rate of Convergence) A sequence $(a_\ell)_{\ell \geq 1}$ of positive numbers is called upper rate of convergence in $L_2^{d\rho_\mathcal{I}}$ norm over the model $\mathcal{P}$, for the sequence of estimated solutions $(f_{\mathbf{z},\lambda_\ell})_{\ell\geq 1}$ using regularization parameters $(\lambda_\ell)_{\ell \geq 0}$ if*

$$\lim_{\tau \to +\infty} \lim_{\ell \to \infty} \sup \sup_{\rho \in \mathcal{P}} \rho^\ell \left(\mathbf{z} : \|f_{\mathbf{z},\lambda_\ell} - f_\rho\|_\rho^2 > \tau a_\ell \right) = 0$$

Definition 5.2. *(Minimax Lower Rate of Convergence) A sequence $(w_\ell)_{\ell \geq 1}$ of positive numbers is called minimax lower rate of convergence in $L_2^{d\rho_\mathcal{I}}$ norm over the model $\mathcal{P}$ if*

$$\lim_{\tau \to 0^+} \lim_{\ell \to \infty} \inf \inf_{f_\mathbf{z}} \sup_{\rho \in \mathcal{P}} \rho^\ell \left(\mathbf{z} : \|f_\mathbf{z} - f_\rho\|_\rho^2 > \tau w_\ell \right) = 1$$

where the infimum is taken over all measurable learning methods with respect to $\mathcal{P}$.

We call such sequences $(w_\ell)_{\ell \geq 1}$ (minimax) lower rates. Every sequence $(\hat{w}_\ell)_{\ell \geq 1}$ which decreases at least with the same speed as $(w_\ell)_{\ell \geq 1}$ is also a lower rate for this set of probability measures and on every larger set of probability measures at least the same lower rate holds. The meaning of a lower rate $(w_\ell)_{\ell \geq 1}$ is that no measurable learning method can achieve a $L_2^{d\rho_\mathcal{I}}(\mathcal{I})$-convergence rate $(a_\ell)_{\ell \geq 1}$ in the sense of Definition 5.1 decreasing faster than $(w_\ell)_{\ell \geq 1}$. In the case where the convergence rate of the sequence coincides with the minimax lower rates, we say that it is optimal in the minimax sense from a nonparametric learning viewpoint.

Setting. Here the hypothesis space considered is the RKHS H_N associated to the Kernel K_N introduced in Prop. 1 where, $N \geq 2$, f_1 a function which admits a Taylor decomposition on $[-1, 1]$ with non-negative coefficients $(b_m)_{m \geq 0}$ and $(f_i)_{i=2}^N$ a sequence of real valuated functions such that $g := f_N \circ \dots . \circ f_2$ admits a Taylor decomposition on $\mathbb{R}$ with non-negative coefficients. In the following, we denote by $T_{\rho_\mathcal{I}}$ the integral operator on $L_2^{d\rho_\mathcal{I}}(\mathcal{I})$ associated with K_N defined as

$$\begin{array}{cccc} T_{\rho_\mathcal{I}} & : & L_2^{d\rho_\mathcal{I}}(\mathcal{I}) & \to & L_2^{d\rho_\mathcal{I}}(\mathcal{I}) \\ & & f & \to & \int_\mathcal{I} K_N(x, \cdot) f(x) d\rho_\mathcal{I}(x) \end{array}$$

Let us now introduce the general assumptions on the class of probability measures considered. Let us denote $dP := \otimes_{i=1}^n d\sigma_{d-1}$ and for $\omega \geq 1$, we denote by $\mathcal{W}_\omega$ the set of all probability measures ν on $\mathcal{I}$ satisfying $\frac{d\nu}{dP} < \omega$. Furthermore, we introduce for a constant $\omega \geq 1 > h > 0$, $\mathcal{W}_{\omega,h} \subset \mathcal{W}_\omega$ the set of probability measures μ on $\mathcal{I}$ which additionally satisfy $\frac{d\nu}{dP} > h$.

Assumptions 5.1. ***Probability measures on** $\mathcal{I} \times \mathcal{Y}$: Let $B, B_\infty, L, \sigma > 0$ be some constants and $0 < \beta \leq 2$ a parameter. We denote by $\mathcal{F}_{B,B_\infty,L,\sigma,\beta}(\mathcal{P})$ the set of all probability measures ρ on $\mathcal{I} \times \mathcal{Y}$ with the following properties*

- *$\rho_\mathcal{I} \in \mathcal{P}$, $\int_{\mathcal{I}\times\mathcal{Y}} y^2 d\rho(x, y) < \infty$ and $\|f_\rho\|^2_{L_\infty^{d\rho_\mathcal{I}}} \leq B_\infty$*
- *there exists $g \in L_2^{d\rho_\mathcal{I}}(\mathcal{I})$ such that $f_\rho = T_{\rho_\mathcal{I}}^{\beta/2} g$ and $\|g\|_\rho^2 \leq B$*
- *there exist $\sigma > 0$ and $L > 0$ such that $\int_\mathcal{Y} |y - f_\rho(x)|^m d\rho(y|x) \leq \frac{1}{2} m! L^{m-2}$*

A sufficient condition for the last assumption is that ρ is concentrated on $\mathcal{I} \times [-M, M]$ for some constant $M > 0$.

In the following we denote $\mathcal{G}_{\omega,\beta} := \mathcal{F}_{B,B_\infty,L,\sigma,\beta}(\mathcal{W}_\omega)$ and $\mathcal{G}_{\omega,h,\beta} := \mathcal{F}_{B,B_\infty,L,\sigma,\beta}(\mathcal{W}_{\omega,h})$.

Remark 1. *Note that we do not make any assumption on the set of distributions related to the eigenvalue decay. Indeed, the control of the eigenvalue decay obtained in Proposition 4.3 allows us to define a non-trivial set of distributions adapted to these kernels.*

The main result of this paper is given in the following theorem. See (Scetbon & Harchaoui, 2020) for details.

Theorem 5.1. *Let us assume there exists $1 > r > 0$ and $c_1 > 0$ a constant such that $(b_m)_{m\geq 0}$ satisfies for all $m \geq 0$ we have $b_m \leq c_1 r^m$. Moreover let us assume that $f_N \circ \ldots \circ f_2$ is a polynomial of degree $D \geq 1$ and let us denote $d^* := \min(D, n)$. Let also $w \geq 1$ and $0 < \beta \leq 2$. Then there exists $A, C > 0$ some constants independent of β such that for any $\rho \in \mathcal{G}_{\omega,\beta}$ and $\tau \geq 1$ we have:*

- *If $\beta > 1$, then for $\lambda_\ell = \frac{1}{\ell^{1/\beta}}$ and $\ell \geq \max\left(e^\beta, \left(\frac{A}{\beta(d-1)d^*}\right)^{\frac{\beta}{\beta-1}} \tau^{\frac{2\beta}{\beta-1}} \log(\ell)^{\frac{(d-1)d^*\beta}{\beta-1}}\right)$, with a ρ^ℓ-probability $\geq 1 - e^{-4\tau}$ it holds*

$$\|f_{H_N,\mathbf{z},\lambda_\ell} - f_\rho\|_\rho^2 \leq 3C\tau^2 \frac{\log(\ell)^{(d-1)d^*}}{\ell}$$

- *If $\beta = 1$, then for $\lambda_\ell = \frac{\log(\ell)^\mu}{\ell}$, $\mu > (d-1)d^* > 0$ and $\ell \geq \max\left(\exp\left((A\tau)^{\frac{1}{\mu-(d-1)d^*}}\right), e^1 \log(\ell)^\mu\right)$, with a ρ^ℓ-probability $\geq 1 - e^{-4\tau}$ it holds*

$$\|f_{H_N,\mathbf{z},\lambda_\ell} - f_\rho\|_\rho^2 \leq 3C\tau^2 \frac{\log(\ell)^\mu}{\ell^\beta}$$

- *If $\beta < 1$, then for $\lambda_\ell = \frac{\log(\ell)^{\frac{(d-1)d^*}{\beta}}}{\ell}$ and $\ell \geq \max\left(\exp\left((A\tau)^{\frac{\beta}{(d-1)d^*(1-\beta)}}\right), e^1 \log(\ell)^{\frac{(d-1)d^*}{\beta}}\right)$, with a ρ^ℓ-probability $\geq 1 - e^{-4\tau}$ it holds*

$$\|f_{H_N,\mathbf{z},\lambda_\ell} - f_\rho\|_\rho^2 \leq 3C\tau^2 \frac{\log(\ell)^{(d-1)d^*}}{\ell^\beta}$$

Remark 2. *It is worth noting that the convergence rates obtained here do not depend on the number of parameters considered in the network which may be much larger than the input dimension. Indeed, here we show that even on the largest possible function space generated by convolutional networks, learning from data can still happen.*

In fact from the above theorem, we can deduce asymptotic upper rate of convergence. Indeed we have

$$\lim_{\tau\to+\infty} \limsup_{\ell\to\infty} \sup_{\rho\in\mathcal{G}_{\omega,\beta}} \rho^\ell\left(\mathbf{z} : \|f_{\mathbf{z},\lambda_\ell} - f_\rho\|_\rho^2 > \tau a_\ell\right) = 0$$

if one of the following conditions hold

- $\beta > 1$, $\lambda_\ell = \frac{1}{\ell^{1/\beta}}$ and $a_\ell = \frac{\log(\ell)^{(d-1)d^*}}{\ell}$
- $\beta = 1$, $\lambda_\ell = \frac{\log(\ell)^\mu}{\ell}$ and $a_\ell = \frac{\log(\ell)^\mu}{\ell}$ for $\mu > (d-1)d^* > 0$
- $\beta < 1$,$\lambda_\ell = \frac{\log(\ell)^{\frac{(d-1)d^*}{\beta}}}{\ell}$ and $a_\ell = \frac{\log(\ell)^{(d-1)d^*}}{\ell^\beta}$

In order to investigate the optimality of the convergence rates, let us take a look at the lower rates.

Theorem 5.2. *Under the exact same assumptions of Theorem 5.1, and if we assume in addition that there exist a constant $0 < c_2 < c_1$ such that for all $m \geq 0$:*

$$c_2 r^m \leq b_m$$

we have that for any $0 < \beta \leq 2$ and $\omega \geq 1 > h > 0$ such that $\mathcal{W}_{\omega,h}$ is not empty

$$\lim_{\tau\to 0^+} \liminf_{\ell\to\infty} \inf_{f_\mathbf{z}} \sup_{\rho\in\mathcal{G}_{\omega,h,\beta}} \rho^\ell\left(\mathbf{z} : \|f_\mathbf{z} - f_\rho\|_\rho^2 > \tau w_\ell\right) = 1$$

where

$$w_\ell = \frac{\log(\ell)^{(d-1)d^*}}{\ell}$$

The infimum is taken over all measurable learning methods with respect to $\mathcal{G}_{\omega,h,\beta}$.

Rate optimality. If the source condition is satisfied with $\beta > 1$, then the convergence rate of the regularized least-squares estimator stated in Theorem 5.1 is optimal in the minimax sense from a nonparametric learning viewpoint (Györfi et al., 2002; Caponnetto & De Vito, 2007; Steinwart & Christmann, 2008).

It is worthwhile to note that the rate is close to the known optimal rate for nonparametric regression with d-dimensional inputs, setting the dimension of the sub-images or patches to d

$$\|f_{\mathbf{z},\lambda_\ell} - f_\rho\|_\rho^2 \leq 3C\tau^2 \frac{\log(\ell)^{d-1}}{\ell}.$$

This connection highlights that the dimension of the sub-images or patches drives the statistical rate of convergence in this regime.

Functional ANOVA. In (Lin, 2000), the author establishes a similar result for a functional ANOVA models assuming that the main effects live in $H^m([0,1])$. Indeed, denoting d^* the highest order of interaction in the model, the regularized least-squares estimator enjoys a near-optimal rate of convergence, within a log factor of the optimal rate of convergence in one dimension

$$\|f_{\mathbf{z},\lambda_\ell} - f_\rho\|_\rho^2 \leq 3C\tau^2 \left(\frac{\log(\ell)^{d^*-1}}{\ell}\right)^{\frac{2m}{2m+1}}.$$

This correspondence brings to light how the construction underlying a convolutional network allows one to overcome the curse of dimensionality. The rates in Theorem 5.1 highlight two important aspects of the behavior of CNNs. First, the highest order of interactions, given by the network depth, controls the statistical performance of such models. If the order is small, we obtain optimal rates which are close to the optimal rate for estimating multivariate functions in d dimensions where d is the patch size. Therefore we obtain learning rates which are almost free dimension.

Second, adding layers makes the eigenvalue decay decrease slower and as soon as $\sigma_N \circ \cdots \circ \sigma_2$ are arbitrary polynomial functions with degrees higher than n, then the optimal rates will be exactly the same as the one obtain for a polynomial function of degree n. There is thus a regime in which adding layers does not affect the convergence rate of convergence, and allows the function space of target functions to grow. Indeed the eigenvalue decay gives a concrete notion of the complexity of the function space considered. Given an eigensystem $(\mu_m)_{m\geq 0}$ and $(e_m)_{m\geq 0}$ of positive eigenvalues and eigenfunctions respectively of the integral operator T_{K_N}, associated with the Kernel K_N, defined on $L_2(\mathcal{I})$, the RKHS H_N associated is defined as

$$H_N = \left\{ f \in L_2(\mathcal{I})\colon f = \sum_{m\geq 0} a_m e_m,\ \left(\frac{a_m}{\sqrt{\mu_m}}\right) \in \ell_2 \right\}$$

endowed with the inner product $\langle f, g\rangle = \sum_{m\geq 0} a_m b_m / \mu_m$. From this definition, we see that, as the rate of decay of the eigenvalues of the integral operator gets slower, the RKHS gets larger. Therefore composing layers allows the function space generated by the network to grow and therefore allows the function space of the target function to grow, while the rates remain the same.

6. Related works

In (Caponnetto & De Vito, 2007), the authors obtained optimal convergence rates of the regularized least-squares estimator for any RKHS yet given a hypothetical set of distributions. Indeed the authors consider a subset of the set of all the distributions for which the eigenvalue decay of the integral operator associated to the kernel is polynomial. This assumption may be too stringent or unsuited for the kernel we consider here. Recall that the eigenvalue decay we are dealing with here is geometric instead.

We show how to control the eigenvalue decay of the integral operator associated to the kernels introduced in Prop. 1 under general assumptions on the activation functions. This control allows us to circumvent abstract assumptions stated in terms of sets of distributions leading to the desired eigenvalue decay as in (Caponnetto & De Vito, 2007). Thus, thanks to the spectral characterization of the kernels we consider, we can actually put forth a non-trivial set of distributions for which the RLS estimator with the corresponding RKHS-s enjoys optimal convergence rates from a nonparametric learning viewpoint. Moreover, this set of distributions is independent of the choice of the RKHS (except for the source condition which can just be fixed). Therefore we are able to compare the convergence rates obtained for the different RKHS-s defined in Prop. 1 on this set of distributions. In particular, we can compare convergence rates depending on different network depths.

In (Bach, 2017), the author considers a single-hidden layer neural network with affine transforms and homogeneous functions acting on vectorial data. In this particular case, the author provides a detailed theoretical analysis of generalization performance. See *e.g.* (Barron, 1994; Anthony & Bartlett, 2009; Mohri et al., 2012) for related classical approaches and (Zhang et al., 2016; 2017) for more recent ones to analyze multi-layer perceptrons.

Recent works (Bartlett et al., 2017; Neyshabur et al., 2018) studied various kinds of statistical bounds for multi-layer perceptrons. In (Bietti & Mairal, 2017), statistical bounds for convolutional kernel networks are presented. These statistical bounds typically depend on the product of spectral norms of matrices stacking layer weights. When put in our context, these bounds do not involve the full eigenspectrum of the integral operator associated with each layer.

We focused here on multi-layer convolutional networks on images. With appropriate changes, a similar analysis can be carried out for signal data, with sub-signals/windows in place of sub-images/patches, and any lattice-structured data (including *e.g.* voxel data).

7. Conclusion.

We have presented an approach to convolutional networks that allowed us to draw connections to nonparametric statistics. In particular, we have brought to light a decomposition akin to a functional ANOVA decomposition that explains how a convolutional network models interactions between sub-images or patches of images. This correspondence allows us to interpret how a convolutional network overcomes the curse of dimensionality when learning from dense images. The extensions of our work beyond least-squares estimators would be an interesting venue for future work.

Acknowledgements. This work was developed during academic year 2018-2019, as M. Scetbon then with ENS Paris-Saclay was visiting Univ. Washington. We acknowledge support from NSF CCF-1740551, NSF DMS-1839371, CIFAR LMB, and faculty research awards.

References

Anthony, M. and Bartlett, P. L. *Neural Network Learning: Theoretical Foundations*. Cambridge UP, New York, NY, USA, 2009.

Bach, F. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

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

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Bietti, A. and Mairal, J. Invariance and stability of deep convolutional representations. In *Advances in Neural Information Processing Systems*, 2017.

Boucheron, S., Bousquet, O., and Lugosi, G. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

Bruna, J. and Mallat, S. Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8): 1872–1886, 2013.

Caponnetto, A. and De Vito, E. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Carmeli, C., Vito, E. D., Toigo, A., and Umanità, V. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 08(01):19–61, 2010.

Efthimiou, C. and Frye, C. *Spherical harmonics in p dimensions*. World Scientific, 2014.

Folland, G. B. *A course in abstract harmonic analysis*. Chapman and Hall/CRC, 2016.

Fukushima, K. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36 (4):193, 1980.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. The MIT Press, 2016.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. *A Distribution-Free Theory of Nonparametric Regression*. Springer series in statistics. Springer, 2002.

LeCun, Y., Bengio, Y., et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

Lin, Y. Tensor product space ANOVA models. *The Annals of Statistics*, 28(3):734–755, 2000.

Mairal, J., Bach, F. R., and Ponce, J. Sparse modeling for image and vision processing. *Found. Trends Comput. Graph. Vis.*, 8(2-3):85–283, 2014a.

Mairal, J., Koniusz, P., Harchaoui, Z., and Schmid, C. Convolutional kernel networks. In *Advances in neural information processing systems*, pp. 2627–2635, 2014b.

Mallat, S. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

Mallat, S. Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065): 20150203, 2016.

Mohri, M., Talwalkar, A., and Rostamizadeh, A. *Foundations of machine learning*. MIT Press Cambridge, MA, 2012.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Oyallon, E., Belilovsky, E., Zagoruyko, S., and Valko, M. Compressing the input for CNN-s with the first-order scattering transform. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Poggio, T. and Anselmi, F. *Visual Cortex and Deep Networks: Learning Invariant Representations*. Computational Neuroscience Series. MIT Press, 2016.

Scetbon, M. and Harchaoui, Z. Harmonic decompositions of convolutional networks. *CoRR*, abs/2003.12756, 2020.

Schölkopf, B. and Smola, A. J. *Learning with Kernels: support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning series. MIT Press, 2002.

Simon, B. *Trace ideals and their applications*. AMS, 2010.

Smale, S. and Zhou, D.-X. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12 (Jul):2389–2410, 2011.

Steinwart, I. and Christmann, A. *Support vector machines*. Springer Science & Business Media, 2008.

Zhang, Y., Lee, J. D., and Jordan, M. I. ℓ_1-regularized neural networks are improperly learnable in polynomial time. In *International Conference on Machine Learning*, 2016.

Zhang, Y., Liang, P., and Wainwright, M. J. Convexified convolutional neural networks. In *International Conference on Machine Learning*, 2017.