Invariance and Stability of Deep Convolutional Representations
Alberto Bietti, Julien Mairal

To cite this version:

HAL Id: hal-01630265
https://inria.hal.science/hal-01630265
Submitted on 7 Nov 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Invariance and Stability of Deep Convolutional Representations

Alberto Bietti
Inria
alberto.bietti@inria.fr

Julien Mairal
Inria
julien.mairal@inria.fr

Abstract

In this paper, we study deep signal representations that are near-invariant to groups of transformations and stable to the action of diffeomorphisms without losing signal information. This is achieved by generalizing the multilayer kernel introduced in the context of convolutional kernel networks and by studying the geometry of the corresponding reproducing kernel Hilbert space. We show that the signal representation is stable, and that models from this functional space, such as a large class of convolutional neural networks, may enjoy the same stability.

1 Introduction

The results achieved by deep neural networks for prediction tasks have been impressive in domains where data is structured and available in large amounts. In particular, convolutional neural networks (CNNs) [14] have shown to model well the local appearance of natural images at multiple scales, while also representing images with some invariance through pooling operations. Yet, the exact nature of this invariance and the characteristics of functional spaces where convolutional neural networks live are poorly understood; overall, these models are sometimes only seen as clever engineering black boxes that have been designed with a lot of insight collected since they were introduced.

Understanding the geometry of these functional spaces is nevertheless a fundamental question. In addition to potentially bringing new intuition about the success of deep networks, it may for instance help solving the issue of regularization, by providing ways to control the variations of prediction functions in a principled manner. Small deformations of natural signals often preserve their main characteristics, such as the class label in a classification task (e.g., the same digit with different handwritings may correspond to the same images up to small deformations), and provide a much richer class of transformations than translations. Representations that are stable to small deformations allow more robust models that may exploit these invariances, which may lead to improved sample complexity. The scattering transform [5, 17] is a recent attempt to characterize convolutional multilayer architectures based on wavelets. The theory provides an elegant characterization of invariance and stability properties of signals represented via the scattering operator, through a notion of Lipschitz stability to the action of diffeomorphisms. Nevertheless, these networks do not involve “learning” in the classical sense since the filters of the networks are pre-defined, and the resulting architecture differs significantly from the most used ones.

In this work, we study these theoretical properties for more standard convolutional architectures from the point of view of positive definite kernels [27]. Specifically, we consider a functional space derived from a kernel for multi-dimensional signals, which admits a multilayer and convolutional structure that generalizes the construction of convolutional kernel networks (CKNs) [15, 16]. We show that this functional space contains a large class of CNNs with smooth homogeneous activation functions in addition to CKNs [15], allowing us to obtain theoretical results for both classes of models.
The main motivation for introducing a kernel framework is to study separately data representation and predictive models. On the one hand, we study the translation-invariance properties of the kernel representation and its stability to the action of diffeomorphisms, obtaining similar guarantees as the scattering transform [17], while preserving signal information. When the kernel is appropriately designed, we also show how to obtain signal representations that are near-invariant to the action of any group of transformations. On the other hand, we show that these stability results can be translated to predictive models by controlling their norm in the functional space. In particular, the RKHS norm controls both stability and generalization, so that stability may lead to improved sample complexity.

Related work. Our work relies on image representations introduced in the context of convolutional kernel networks [15,16], which yield a sequence of spatial maps similar to traditional CNNs, but each point on the maps is possibly infinite-dimensional and lives in a reproducing kernel Hilbert space (RKHS). The extension to signals with $d$ spatial dimensions is straightforward. Since computing the corresponding Gram matrix as in classical kernel machines is computationally impractical, CKNs provide an approximation scheme consisting of learning finite-dimensional subspaces of each RKHS’s layer, where the data is projected, see [15]. The resulting architecture of CKNs resembles traditional CNNs with a subspace learning interpretation and different unsupervised learning principles.

Another major source of inspiration is the study of group-invariance and stability to the action of diffeomorphisms of scattering networks [17], which introduced the main formalism and several proof techniques from harmonic analysis that were keys to our results. Our main effort was to extend them to more general CNN architectures and to the kernel framework. Invariance to groups of transformations was also studied for more classical convolutional neural networks from methodological and empirical points of view [6,9], and for shallow learned representations [1] or kernel methods [13,19,22]. Note also that other techniques combining deep neural networks and kernels have been introduced. Early multilayer kernel machines appear for instance in [7,26]. Shallow kernels for images modelling local regions were also proposed in [25], and a multilayer construction was proposed in [4]. More recently, different models based on kernels are introduced in [2,10,18] to gain some theoretical insight about classical multilayer neural networks, while kernels are used to define convex models for two-layer neural networks in [26]. Finally, we note that Lipschitz stability of deep models to additive perturbations was found to be important to get robustness to adversarial examples [8]. Our results show that convolutional kernel networks already enjoy such a property.

Notation and basic mathematical tools. A positive definite kernel $K$ that operates on a set $\mathcal{X}$ implicitly defines a reproducing kernel Hilbert space $\mathcal{H}$ of functions from $\mathcal{X}$ to $\mathbb{R}$, along with a mapping $\varphi : \mathcal{X} \to \mathcal{H}$. A predictive model associates to every point $z$ in $\mathcal{X}$ a label in $\mathbb{R}$; it consists of a linear function $f$ in $\mathcal{H}$ such that $f(z) = \langle f, \varphi(z) \rangle_{\mathcal{H}}$, where $\varphi(z)$ is the data representation. Given now two points $z, z' \in \mathcal{X}$, Cauchy-Schwarz’s inequality allows us to control the variation of the predictive model $f$ according to the geometry induced by the Hilbert norm $\| \cdot \|_{\mathcal{H}}$:

$$|f(z) - f(z')| \leq \|f\|_{\mathcal{H}} \|\varphi(z) - \varphi(z')\|_{\mathcal{H}}.$$  

This property implies that two points $z$ and $z'$ that are close to each other according to the RKHS norm should lead to similar predictions, when the model $f$ has reasonably small norm in $\mathcal{H}$.

Then, we consider notation from signal processing similar to [17]. We call a signal $x$ a function in $L^2(\Omega, \mathcal{H})$, where $\Omega$ is a subset of $\mathbb{R}^d$ representing spatial coordinates, and $\mathcal{H}$ is a Hilbert space, when $\|x\|^2_\Omega := \int_{\Omega} \|x(u)\|^2_{\mathcal{H}} du < \infty$, where $du$ is the Lebesgue measure on $\mathbb{R}^d$. Given a linear operator $T : L^2(\Omega, \mathcal{H}) \to L^2(\Omega, \mathcal{H}')$, the operator norm is defined as $\|T\|_{L^2(\Omega, \mathcal{H}) \to L^2(\Omega, \mathcal{H}')} := \sup\{\|T(x)\|_{\Omega} : \|x\|_{L^2(\Omega, \mathcal{H})} \leq 1\}$. For the sake of clarity, we drop norm subscripts, from now on, using the notation $\| \cdot \|$ for Hilbert space norms, $L^2$ norms, and $L^2 \to L^2$ operator norms, while $| \cdot |$ denotes the Euclidean norm on $\mathbb{R}^d$. Some useful mathematical tools are also presented in Appendix A.

2 Construction of the Multilayer Convolutional Kernel

We now present the multilayer convolutional kernel, which operates on signals with $d$ spatial dimensions. The construction follows closely that of convolutional kernel networks [15] but generalizes it to input signals defined on the continuous domain $\Omega = \mathbb{R}^d$ (which does not prevent signals to have compact support), as done by Mallat [17] for analyzing the properties of the scattering transform; the issue of discretization where $\Omega$ is a discrete grid is addressed in Section 2.4.
Applying Schur’s test to the integral operator which ensures that \( \|A_k\| \leq 1 \). Thus, \( \|x_k\| \leq \|M_k P_k x_{k-1}\| \) and \( x_k \in L^2(\Omega, \mathcal{H}_k) \). Note that a similar pooling operator is used in the scattering representation [5, 17], though in a different way which does not affect subsequent layers.
Multilayer construction. Finally, we obtain a multilayer representation by composing multiple times the previous operators. In order to increase invariance with each layer, the size of the patch $S_k$ and pooling scale $\sigma_k$ typically grow exponentially with $k$, with $\sigma_k$ and $\sup_{c \in S_k} |c|$ of the same order. With $n$ layers, the final representation is given by the feature map

$$\Phi_n(x_0) := x_n = A_n M_n P_n A_{n-1} M_{n-1} P_{n-1} \cdots A_1 M_1 P_1 x_0 \in L^2(\Omega, H_n).$$

Then, we can define a kernel $K_n$ on two signals $x_0$ and $x'_0$ by $K_n(x_0, x'_0) := \langle \Phi_n(x_0), \Phi_n(x'_0) \rangle$, whose RKHS $H_{K_n}$ contains all functions of the form $f(x_0) = \langle w, \Phi_n(x_0) \rangle$ with $w \in L^2(\Omega, H_n)$.

The following lemma shows that this representation preserves all information about the signal at each layer, and each feature map $x_k$ can be sampled on a discrete set with no loss of information. This suggests a natural approach for discretization which we discuss next. For space limitation reasons, all proofs in this paper are relegated to Appendix C.

Lemma 1 (Signal preservation). Assume that $H_k$ contains linear functions $\langle w, \cdot \rangle$ with $w \in P_k$ (this is true for all kernels $K_k$ described in Appendix B), then the signal $x_{k-1}$ can be recovered from a sampling of $x_k = A_k M_k \tilde{P}_k x_{k-1}$ at discrete locations as soon as the union of patches centered at these points covers all of $\Omega$. It follows that $x_k$ can be reconstructed from such a sampling.

2.1 From Theory to Practice: Discretization and Signal Preservation

The previous construction defines a kernel representation for general signals in $L^2(\Omega, H_0)$, which is an abstract object defined for theoretical purposes, as often done in signal processing [17]. In practice, signals are discrete, and it is thus important to discuss the problem of discretization, as done in [15]. For clarity, we limit the presentation to 1-dimensional signals ($\Omega = \mathbb{R}^d$ with $d = 1$), but the arguments can easily be extended to higher dimensions when using box-shaped patches. Notation from the previous section is preserved, but we add a bar on top of all discrete analogues of their discrete counterparts, e.g., $\bar{x}_k$ is a discrete feature map in $\ell^2(\mathbb{Z}, \bar{H}_k)$ for some RKHS $\bar{H}_k$.

Input signals $x_0$ and $\bar{x}_0$. Discrete signals acquired by a physical device are often seen as local integrators of signals defined on a continuous domain (e.g., sensors from digital cameras integrate the pointwise distribution of photons that hit a sensor in a spatial window). Let us then consider a signal $x_0$ in $L^2(\Omega, H_0)$ and $s_0$ a sampling interval. By defining $\bar{x}_0$ in $\ell^2(\mathbb{Z}, H_0)$ such that $\bar{x}_0[n] = x_0(ns_0)$ for all $n$ in $\mathbb{Z}$, it is thus natural to assume that $x_0 = A_0 x$, where $A_0$ is a pooling operator (local integrator) applied to an original signal $x$. The role of $A_0$ is to prevent aliasing and reduce high frequencies; typically, the scale $\sigma_0$ of $A_0$ should be of the same magnitude as $s_0$, which we choose to be $s_0 = 1$ in the following, without loss of generality. This natural assumption will be kept later in the analysis.

Multilayer construction. We now want to build discrete feature maps $\bar{x}_k$ in $\ell^2(\mathbb{Z}, \bar{H}_k)$ at each layer $k$ involving subsampling with a factor $s_k$ w.r.t. $\bar{x}_{k-1}$. We now define the discrete analogues of the operators $P_k$ (patch extraction), $M_k$ (kernel mapping), and $A_k$ (pooling) as follows: for $n \in \mathbb{Z}$,

$$\bar{P}_k \bar{x}_{k-1}[n] := \sigma_k^{-1/2}(\bar{x}_{k-1}[n], \bar{x}_{k-1}[n + 1], \ldots, \bar{x}_{k-1}[n + e_k - 1]) \in \bar{P}_k := \bar{H}_k^{e_k-1}$$

$$M_k \bar{P}_k \bar{x}_{k-1}[n] := \phi_k(\bar{P}_k \bar{x}_{k-1}[n]) \in \bar{H}_k$$

$$\bar{x}_k[n] = \bar{A}_k M_k \bar{P}_k \bar{x}_{k-1}[n] := s_k^{1/2} \sum_{m \in \mathbb{Z}} h_k[ns_k - m] M_k \bar{P}_k \bar{x}_{k-1}[m] = \bar{h}_k * M_k \bar{P}_k \bar{x}_{k-1}[ns_k] \in \bar{H}_k,$$

where (i) $\bar{P}_k$ extracts a patch of size $e_k$ starting at position $n$ in $\bar{x}_{k-1}[n]$ (defining a patch centered at $n$ is also possible), which lives in the Hilbert space $\bar{P}_k$ defined as the direct sum of $e_k$ times $\bar{H}_{k-1}$; (ii) $M_k$ is a kernel mapping identical to the continuous case, which preserves the norm, like $M_k$; (iii) $\bar{A}_k$ performs a convolution with a Gaussian filter and a subsampling operation with factor $s_k$.

The next lemma shows that under mild assumptions, this construction preserves signal information.

Lemma 2 (Signal recovery with subsampling). Assume that $\bar{H}_k$ contains the linear functions $\langle w, \cdot \rangle$ for all $w \in \bar{P}_k$ and that $e_k \geq s_k$. Then, $\bar{x}_k$ can be recovered from $\bar{x}_{k-1}$.

We note that this result relies on recovery by deconvolution of a pooling convolution with filter $\bar{h}_k$, which is stable when its scale parameter, typically of order $s_k$ to prevent anti-aliasing, is small enough. This suggests using small values for $e_k$, $s_k$, as in typical recent convolutional architectures [30].
2.2 From Theory to Practice: Kernel Approximation and Convolutional Kernel Networks

Besides discretization, two modifications are required to use the image representation we have described in practice. The first one consists of using feature maps with finite spatial support, which introduces border effects that we did not study, but which are negligible when dealing with large realistic images. The second one requires finite-dimensional approximation of the kernel maps, leading to the convolutional kernel network model of [15]. Typically, each RKHS’s mapping is approximated by performing a projection onto a subspace of finite dimension, a classical approach to make kernel methods work at large scale [12, 31, 34]. One advantage is its compatibility with the RKHSs (meaning that the approximations live in the respective RKHSs), and the stability results we will present next are preserved thanks to the non-expansiveness of the projection.

It is then be possible to derive theoretical results for the CKN model, which appears as a natural implementation of the kernel constructed previously; yet, we will also show in Section 5 that the results apply more broadly to CNNs that are contained in the functional space associated to the kernel.

3 Stability to Deformations and Translation Invariance

In this section, we study the translation-invariance and the stability of the kernel representation described in Section 2 for continuous signals under the action of diffeomorphisms. We use a similar characterization of stability to the one introduced by Mallat [17]: for a $C^1$-diffeomorphism $\tau : \Omega \rightarrow \Omega$, let $L_\tau$ denote the linear operator defined by $L_\tau x(u) = x(u - \tau(u))$, the representation $\Phi(\cdot)$ is stable under the action of diffeomorphisms if there exist two constants $C_1$ and $C_2$ such that

$$\|\Phi(L_\tau x) - \Phi(x)\| \leq (C_1\|\nabla\tau\|_\infty + C_2\|\tau\|_\infty)\|x\|,$$

(4)

where $\nabla\tau$ is the Jacobian of $\tau$, $\|\nabla\tau\|_\infty := \sup_{u \in \Omega} \|\nabla\tau(u)\|$, and $\|\tau\|_\infty := \sup_{u \in \Omega} \|\tau(u)\|$. As in [17], our results will assume the regularity condition $\|\nabla\tau\|_\infty < 1/2$. In order to have a translation-invariant representation, we want $C_2$ to be small (a translation is a diffeomorphism with $\nabla\tau = 0$), and indeed we will show that $C_2$ is proportional to $1/\sigma_n$, where $\sigma_n$ is the scale of the last pooling layer, which typically increases exponentially with the number of layers $n$.

Note that unlike the scattering transform [17], we do not have a representation that preserves the norm, i.e., such that $\|\Phi(x)\| = \|x\|$. While the patch extraction $P_k$ and kernel mapping $M_k$ operators do preserve the norm, the pooling operators $A_k$ may remove (or significantly reduce) frequencies from the signal that are larger than $1/\sigma_k$. Yet, natural signals such as natural images often have high energy in the low-frequency domain (the power spectra of natural images is often considered to have a polynomial decay in $1/f^2$, where $f$ is the signal frequency [33]). For such classes of signals, a large fraction of the signal energy will be preserved by the pooling operator. In particular, with some additional assumptions on the kernels $K_k$, it is possible to show [3]:

$$\|\Phi(x)\| \geq \|A_n \cdots A_0 x\|.$$

Additionally, when using a Gaussian kernel mapping $\varphi_{n+1}$ on top of the last feature map as a prediction layer instead of a linear layer, the final representation $\Phi_f(x) := \varphi_{n+1}(\Phi_n(A_0 x))$ preserves stability and always has unit norm (see the extended version of the paper [3] for details). This suggests that norm preservation may be a less relevant concern in our kernel setting.

3.1 Stability Results

In order to study the stability of the representation [3], we assume that the input signal $x_0$ may be written as $x_0 = A_0 x$, where $A_0$ is an initial pooling operator at scale $\sigma_0$, which allows us to control the high frequencies of the signal in the first layer. As discussed previously in Section 2.1, this assumption is natural and compatible with any physical acquisition device. Note that $\sigma_0$ can be taken arbitrarily small, making the operator $A_0$ arbitrarily close to the identity, so that this assumption does not limit the generality of our results. Moreover, we make the following assumptions for each layer $k$:
Appendix B exposes a large class of homogeneous kernels that satisfy assumptions (A1-2). Let

\[ \sum_{k=1}^{n} ||P_k A_{k-1} - L_t|| + ||A_n, L_t|| + ||L_{\tau} A_n - A_n|| \leq ||x|| \]  

(5)

In the case of a translation \( L_{\tau} x(u) = L_c x(u) = x(u - c) \), it is easy to see that pooling and patch extraction operators commute with \( L_c \) (this is also known as covariance or equivariance to translations), so that we are left with the term \( ||L_{\tau}A_n - A_n|| \), which should control translation invariance. For general diffeomorphisms \( \tau \), we no longer have exact covariance, but we show below that commutators are stable to \( \tau \), in the sense that \( ||P_k A_{k-1}, L_{\tau}|| \) is controlled by \( ||\nabla \tau||_\infty \), while \( ||L_{\tau} A_n - A_n|| \) is controlled by \( ||\tau||_\infty \) and decays with the pooling size \( \sigma_n \).

**Bound on** \( ||P_k A_{k-1}, L_{\tau}|| \). We begin by noting that \( P_k z \) can be identified with \( (L_c z)_{c \in S_k} \) isometrically for all \( z \) in \( L^2(\Omega, \mathcal{H}_k) \), since \( ||P_k z|| = \int_{S_k} ||L_c z||^2 d\nu_k(c) \) by Fubini’s theorem. Then,

\[ ||P_k A_{k-1} L_{\tau} z - L_{\tau} P_k A_{k-1} z||^2 = \int_{S_k} ||L_c A_{k-1} L_{\tau} z - L_{\tau} L_c A_{k-1} z||^2 d\nu_k(c) \leq \sup_{c \in S_k} ||L_c A_{k-1} L_{\tau} x - L_{\tau} L_c A_{k-1} z||^2, \]

so that \( ||P_k A_{k-1}, L_{\tau}|| \leq \sup_{c \in S_k} ||L_c A_{k-1}, L_{\tau}|| \). The following result lets us bound \( ||L_{\tau} A_n - A_n|| \) when \( |c| \leq \kappa \sigma_{k-1} \), which is satisfied under assumption (A3).

**Lemma 4.** Let \( A_n \) be the pooling operator with kernel \( h_n(u) = \sigma^{-d} h(u/\sigma) \). If \( ||\nabla \tau||_\infty \leq 1/2 \), there exists a constant \( C_1 \) such that for any \( \sigma \) and \( |c| \leq \kappa \sigma \), we have

\[ ||L_{\tau} A_n, L_{\tau}|| \leq C_1 ||\nabla \tau||_\infty, \]

where \( C_1 \) depends only on \( h \) and \( \kappa \).

A similar result is obtained in Mallat [17] Lemma E.1 for commutators of the form \( [A_{\tau}, L_{\tau}] \), but we extend it to handle integral operators \( L_c A_{\sigma} \) with a shifted kernel. The proof (given in Appendix C.4) relies on the fact that \( [L_c A_{\sigma}, L_{\tau}] \) is an integral operator in order to bound its norm via Schur’s test. Note that \( \kappa \) can be made larger, at the cost of an increase of the constant \( C_1 \) of the order \( \kappa^{-d+1} \).

**Bound on** \( ||L_{\tau} A_n - A_n|| \). We bound the operator norm \( ||L_{\tau} A_n - A_n|| \) in terms of \( ||\tau||_\infty \) using the following result due to Mallat [17] Lemma 2.11, with \( \sigma = \sigma_n \):

**Lemma 5.** If \( ||\nabla \tau||_\infty \leq 1/2 \), we have

\[ ||L_{\tau} A_n - A_n|| \leq \frac{C_2}{\sigma} ||\tau||_\infty, \]

with \( C_2 = 2^d ||\nabla h||_1 \).

Combining Proposition 3 with Lemmas 4 and 5, we immediately obtain the following result.

**Theorem 6.** Let \( \Phi(x) \) be a representation given by \( \Phi(x) = \Phi_n(A_0 x) \) and assume (A1-3). If \( ||\nabla \tau||_\infty \leq 1/2 \), we have

\[ ||\Phi(L_{\tau} x) - \Phi(x)|| \leq \left( C_1 (1 + n) ||\nabla \tau||_\infty + \frac{C_2}{\sigma_n} ||\tau||_\infty \right) ||x||. \]

(7)
We assume the initial signal $x_h$ where we are interested in defining a layer—that is, a succession of patch extraction, kernel mapping, and pooling.

Another approximation performed in the CKN model of [15] consists of adding projection steps on $\Phi(x)$, as in the analysis of the scattering transform of [17], we have characterized the stability and shift-invariance of the data representation for continuous signals, in order to give some intuition about the properties of the corresponding discrete representation, which we have described in Section 2.1.

As in the analysis of the scattering transform of [17], we have characterized the stability and shift-invariance of the data representation for continuous signals, in order to give some intuition about the properties of the corresponding discrete representation, which we have described in Section 2.1.

Another approximation performed in the CKN model of [15] consists of adding projection steps on finite-dimensional subspaces of the RKHS’s layers, as discussed in Section 2.2. Interestingly, the stability properties we have obtained previously are compatible with these steps. We may indeed redefine the operator $M_k$ as the pointwise operation such that $M_k z(u) = \Pi_k \phi_k(z(u))$ for any map $z$ in $L^2(\Omega, P_k)$, instead of $M_k z(u) = \phi_k(z(u))$; $\Pi_k : \mathcal{H} \rightarrow F_k$ is here a projection operator onto a linear subspace. Then, $M_k$ does not necessarily preserve the norm anymore, but $\|M_k z\| \leq \|z\|$, with a loss of information corresponding to the quality of approximation of the kernel $K_k$ on the points $z(u)$.

On the other hand, the non-expansiveness of $M_k$ is satisfied thanks to the non-expansiveness of the projection. Additionally, the CKN construction provides a finite-dimensional representation at each layer, which preserves the norm structure of the original Hilbert spaces isometrically. In summary, it is possible to show that the conclusions of Theorem 6 remain valid for this tractable CKN representation, but we lose signal information in the process. The stability of the predictions can then be controlled through the norm of the last (linear) layer, which is typically used as a regularizer [15].

4 Global Invariance to Group Actions

In Section 3, we have seen how the kernel representation of Section 2 creates invariance to translations by commuting with the action of translations at intermediate layers, and how the last pooling layer on the translation group governs the final level of invariance. It is often useful to encode invariances to different groups of transformations, such as rotations or reflections (see, e.g., [9, 17, 22, 29]). Here, we show how this can be achieved by defining adapted patch extraction and pooling operators that commute with the action of a transformation group $G$ (this is known as group covariance or equivariance). We assume that $G$ is locally compact, so that we can define a left-invariant Haar measure $\mu$—that is, a measure on $G$ that satisfies $\mu(gs) = \mu(S)$ for any Borel set $S \subset G$ and $g \in G$. We assume the initial signal $x(u)$ is defined on $G$, and we define subsequent feature maps on the same domain. The action of an element $g \in G$ is denoted by $L_g$, where $L_g x(u) = x(g^{-1}u)$.

Patch extraction. We define patch extraction as follows

$$Px(u) = (x(uv))_{v \in \mathcal{S}} \text{ for all } u \in G,$$

where $S \subset G$ is a patch centered at the identity. $P$ commutes with $L_g$ since

$$PL_g x(u) = (L_g x(uv))_{v \in \mathcal{S}} = (x(g^{-1}uv))_{v \in \mathcal{S}} = Px(g^{-1}u) = L_g Px(u).$$

Kernel mapping. The pointwise operator $M$ is defined as in Section 2 and thus commutes with $L_g$.

Pooling. The pooling operator on the group $G$ is defined in a similar fashion as [22] by

$$Ax(u) = \int_G x(uv)h(v)d\mu(v) = \int_G x(v)h(u^{-1}v)d\mu(v),$$

where $h$ is a pooling filter typically localized around the identity element. It is easy to see from the first expression of $Ax(u)$ that $AL_g Ax(u) = L_g Ax(u)$, making the pooling operator $G$-equivariant.
In our analysis of stability in Section 3, we saw that inner pooling layers are useful to guarantee stability to local deformations, while global invariance is achieved mainly through the last pooling layer. In some cases, one only needs stability to a subgroup of $G$, while achieving global invariance to the whole group, e.g., in the roto-translation group $\mathbb{R}^2$, one might want invariance to a global rotation but stability to local translations. Then, one can perform pooling just on the subgroup to stabilize (e.g., translations) in intermediate layers, while pooling on the entire group at the last layer to achieve the global group invariance.

5 Link with Convolutional Neural Networks

In this section, we study the connection between the kernel representation defined in Section 2 and CNNs. Specifically, we show that the RKHS $\mathcal{H}_{K,n}$ obtained from our kernel construction contains a set of CNNs on continuous domains with certain types of smooth homogeneous activations. An important consequence is that the stability results of previous sections apply to this class of CNNs.

**CNN maps construction.** We now define a CNN function $f_\sigma$ that takes as input an image $x_0$ in $L^2(\Omega, \mathbb{R}^{p_0})$ with $p_0$ channels, and builds a sequence of feature maps, represented at layer $k$ as a function $z_k$ in $L^2(\Omega, \mathbb{R}^{p_k})$ with $p_k$ channels; it performs linear convolutions with a set of filters $(w_i^k)_i=1,...,p_k$, followed by a pointwise activation function $\sigma$ to obtain intermediate feature maps $\tilde{z}_k$, then applies a linear pooling filter and repeats the same operations at each layer. Note that here, each $w_i^k$ is in $L^2(S_k, \mathbb{R}^{p_{k-1}})$, with channels denoted by $w_i^k \in L^2(S_k, \mathbb{R})$. Formally, the intermediate map $\tilde{z}_k$ in $L^2(\Omega, \mathbb{R}^{p_k})$ is obtained for $k \geq 1$ by

$$\tilde{z}_k^i(u) = n_k(u)\sigma((w_i^k, P_k z_{k-1}(u))/n_k(u)),$$

where $\tilde{z}_k(u) = (\tilde{z}_k^1(u), \ldots, \tilde{z}_k^{p_k}(u))$ in $\mathbb{R}^{p_k}$, and $P_k$ is the patch extraction operator, which operates here on finite-dimensional maps. The activation involves a pointwise non-linearity $\sigma$ along with a quantity $n_k(u)$ that is independent of the filters and that will be made explicit in the sequel. Finally, the map $z_k$ is obtained by using a pooling operator as in Section 2, with $z_k = A_k \tilde{z}_k$, and $z_0 = x_0$.

**Homogeneous activations.** The choice of non-linearity $\sigma$ relies on Lemma B.2 of the appendix, which shows that for many choices of smooth functions $\sigma$, the RKHS $\mathcal{H}_k$ defined in Section 2 contains the linear functions $z \mapsto \|z\|\sigma(\langle g, z \rangle/\|z\|)$ for all $g$ in $P_k$. While this homogenization involving the quantities $\|z\|$ is not standard in classical CNNs, we note that (i) the most successful activation function, namely rectified linear units, is homogeneous—that is, $\text{relu}(\langle g, z \rangle) = \|z\|\text{relu}(\langle g, z \rangle/\|z\|)$; (ii) while relu is nonsmooth and thus not in our RKHSs, there exists a smoothed variant that satisfies the conditions of Lemma B.2 for useful kernels. As noticed in [35][36], this is for instance the case for the inverse polynomial kernel described in Appendix B. In Figure 2, we plot and compare these different variants of relu. Then, we may now define the quantities $n_k(u) := \|P_k x_{k-1}(u)\|$ in (8), which are due to the homogenization, and which are independent of the filters $w_i^k$.

**Classification layer.** The final CNN prediction function $f_\sigma$ is given by inner products with the feature maps of the last layer:

$$f_\sigma(x_0) = \langle w_{n+1}, z_n \rangle,$$

with parameters $w_{n+1}$ in $L^2(\Omega, \mathbb{R}^{p_n})$. The next result shows that for appropriate $\sigma$, the function $f_\sigma$ is in $\mathcal{H}_{K,n}$. The construction of this function in the RKHS and the proof are given in Appendix D. We note that a similar construction for fully connected networks with constraints on weights and inputs was given in [35].

**Proposition 7** (CNNs and RKHSs). Assume the activation $\sigma$ satisfies $C_\sigma(a) < \infty$ for all $a \geq 0$, where $C_\sigma$ is defined for a given kernel in Lemma B.2. Then the CNN function $f_\sigma$ defined above is in the RKHS $\mathcal{H}_{K,n}$, with norm

$$\|f_\sigma\|^2 \leq p_n \sum_{i=1}^{p_n} \|w_{n+1}^i\|_2^2 B_{n,i},$$

where $B_{n,i}$ is defined recursively by $B_{1,i} = C_\sigma^2(\|w_1^i\|_2^2)$ and $B_{k,i} = C_\sigma^2 \left( p_{k-1} \sum_{j=1}^{p_{k-1}} \|w_k^j\|_2^2 B_{k-1,j} \right)$.

The results of this section imply that our study of the geometry of the kernel representations, and in particular the stability and invariance properties of Section 3, apply to the generic CNNs defined
Figure 2: Comparison of one-dimensional functions obtained with relu and smoothed relu (sReLU) activations. (Left) non-homogeneous setting of \([35, 36]\). (Right) our homogeneous setting, for different values of the parameter \(w\). Note that for \(w \geq 0.5\), sReLU and ReLU are indistinguishable.

above, thanks to the Lipschitz smoothness relation \([1]\). The smoothness is then controlled by the
RKHS norm of these functions, which sheds light on the links between generalization and stability. In particular, functions with low RKHS norm (a.k.a. “large margin”) are known to generalize better to unseen data (see, e.g., the notion of margin bounds for SVMs \([27, 28]\)). This implies, for instance, that generalization is harder if the task requires classifying two slightly deformed images with different labels, since this requires a function with large RKHS norm according to our stability analysis. In contrast, if a stable function (i.e., with small RKHS norm) is sufficient to do well on a training set, learning becomes “easier” and few samples may be enough for good generalization.

Acknowledgements

This work was supported by a grant from ANR (MACARON project under grant number ANR-14-CE23-0003-01), by the ERC grant number 714381 (SOLARIS project), and by the MSR-Inria joint center.

References


10


A Useful Mathematical Tools

In this section, we present preliminary mathematical tools that are used in our analysis.

Harmonic analysis. We recall a classical result from harmonic analysis (see, e.g., [22]), which was used many times in [17] to prove the stability of the scattering transform to the action of diffeomorphisms.

Lemma A.1 (Schur’s test). Let $\mathcal{H}$ be a Hilbert space and $\Omega$ a subset of $\mathbb{R}^d$. Consider $T$ an integral operator with kernel $k : \Omega \times \Omega \to \mathbb{R}$, meaning that for all $u \in \Omega$ and $x$ in $L^2(\Omega, \mathcal{H})$,

$$Tx(u) = \int_\Omega k(u, v)x(v)dv,$$

where the integral is a Bochner integral (see, [17, 20]) when $\mathcal{H}$ is infinite-dimensional. If

$$\forall u \in \Omega, \quad \int |k(u, v)|dv \leq C \quad \text{and} \quad \forall v \in \Omega, \quad \int |k(u, v)|du \leq C,$$

for some constant $C$, then, $Tx$ is always in $L^2(\Omega, \mathcal{H})$ for all $x$ in $L^2(\Omega, \mathcal{H})$ and we have $\|T\| \leq C$.

Note that while the proofs of the lemma above are typically given for real-valued functions in $L^2(\Omega, \mathbb{R})$, the result can easily be extended to Hilbert space-valued functions $x$ in $L^2(\Omega, \mathcal{H})$. In order to prove this, we consider the integral operator $|T|$ with kernel $|k|$ that operates on $L^2(\Omega, \mathbb{R})$, meaning that $|T|$ is defined as in (9) by replacing $|k(u, v)|$ by the absolute value $|k(u, v)|$. Then, consider $x$ in $L^2(\Omega, \mathcal{H})$ and use the triangle inequality property of Bochner integrals:

$$\|Tx\|^2 = \int_\Omega \|Tx(u)\|^2du \leq \int_\Omega \left(\int_\Omega |k(u, v)||x(v)||dv\right)^2du = \|[T]|x\|^2,$$

where the function $|x|$ is such that $|x|(u) = \|x(u)\|$ and thus $|x|$ is in $L^2(\Omega, \mathbb{R}_+)$. We may now apply Schur’s test to the operator $|T|$ for real-valued functions, which gives $\|[T]\| \leq C$. Then, noting that $\|[x]\| = \|x\|$, we conclude with the inequality $\|Tx\|^2 \leq \|[T]|x\|^2 \leq \|[T]\|^2\|x\|^2 \leq C^2\|x\|^2$.

The following lemma shows that the pooling operators $A_k$ defined in Section 2 are non-expansive.

Lemma A.2 (Non-expansiveness of pooling operators). If $h(u) := (2\pi)^{-d/2}\exp(-|u|^2/2)$, then the pooling operator $A_\sigma$ defined for any $\sigma > 0$ by

$$A_\sigma x(u) = \int_{\mathbb{R}^d} \sigma^{-d}h\left(\frac{u-v}{\sigma}\right)x(v)dv,$$

has operator norm $\|A_\sigma\| \leq 1$.

Proof. $A_\sigma$ is an integral operator with kernel $k(u, v) := \sigma^{-d}h((u-v)/\sigma)$. By change of variables, we have

$$\int_{\mathbb{R}^d} |k(u, v)|dv = \int_{\mathbb{R}^d} |k(u, v)|dv = \int_{\mathbb{R}^d} h(u)du = 1,$$

since $h$ is a standard Gaussian and thus integrates to 1. The result follows from Schur’s test.

Kernel methods. We now recall a classical result thatcharacterizes the reproducing kernel Hilbert space (RKHS) of functions defined from explicit Hilbert space mappings (see, e.g., [23, §2.1]).

Theorem A.1. Let $\psi : \mathcal{X} \to H$ be a feature map to a Hilbert space $H$, and let $K(z, z') := \langle \psi(z), \psi(z') \rangle_H$ for $z, z' \in \mathcal{X}$. Let $\mathcal{H}$ be the Hilbert space defined by

$$\mathcal{H} := \{ f = \langle w, \psi(\cdot) \rangle_H, w \in H \}$$

$$\|f\|^2_\mathcal{H} := \inf_{w \in H} \{ \|w\|^2_H, f = \langle w, \psi(\cdot) \rangle_H \}.$$

Then $\mathcal{H}$ is the RKHS associated to kernel $K$.

A consequence of this result is that RKHS of the kernel $K_n(x, x') = \langle \Phi_n(x), \Phi_n(x') \rangle$ defined from the last layer representations $\Phi_n(x) \in L^2(\Omega, \mathcal{H}_n)$ introduced in (3) contains functions of the form $f : x \mapsto \langle w, \Phi(x) \rangle$ with $w \in L^2(\Omega, \mathcal{H}_n)$, and the RKHS norm of such a function satisfies $\|f\| \leq \|w\|_{L^2(\Omega, \mathcal{H}_n)}$.
B Choices of Basic Kernels

In this section, we characterize the basic kernels $K_k$ that may be used in the construction of the multilayer convolutional kernel described in Section 2. We recall here the shape of these kernels, which operate on a given Hilbert space $H_0$. We consider, for $z, z' \in H_0$,

$$K(z, z') = \|z\| \|z'\| \kappa \left( \frac{\langle z, z' \rangle}{\|z\| \|z'\|} \right),$$  \hspace{1cm} (10)

which is positive definite when $\kappa$ admits a Maclaurin expansion with only non-negative coefficients $\sum_{j=0}^{\infty} b_j u^j$ with $b_j \geq 0$ for all $u$ in $[-1, 1]$. Let $\varphi(\cdot)$ denote the kernel mapping associated to $K$, so that $K(z, z') = \langle \varphi(z), \varphi(z') \rangle$.

For our stability analysis, we desire the following properties:

- **norm preservation**: $\|\varphi(z)\| = \|z\|$; this is ensured by the condition $\kappa(1) = 1$.
- **non-expansive mapping**: $\|\varphi(z) - \varphi(z')\| \leq \|z - z'\|$.

Even though our stability results make the non-expansive assumption, they can be easily extended to Lipschitz continuous mappings. Then, the Lipschitz constants would appear in the upper-bounds from Section 3, and the stability constants would depend exponentially in the number of layers, which we would like to avoid. Below, we give a simple lemma to characterize kernels with the non-expansiveness property, and show that it applies to a large class of useful examples.

**Lemma B.1** (Non-expansive homogeneous kernel mappings). Let $K$ be a kernel of the form \[(10). \] If $\kappa$ is convex, $\kappa(1) = 1$, and $0 \leq \kappa'(1) \leq 1$, where $\kappa'$ denotes the first derivative of $\kappa$, then the kernel mapping is non-expansive.

**Proof.** First, we notice that

$$\|\varphi(z) - \varphi(z')\|^2 = K(z, z) + K(z', z') - 2K(z, z') = \|z\|^2 + \|z'\|^2 - 2\|z\|\|z'\|\kappa(u),$$

with $u = \langle z, z' \rangle / (\|z\|\|z'\|)$. Since $\kappa$ is convex, we also have $\kappa(u) \geq \kappa(1) + \kappa'(1)(u - 1) = 1 + \kappa'(1)(u - 1)$, and

$$\|\varphi(z) - \varphi(z')\|^2 \leq \|z\|^2 + \|z'\|^2 - 2\|z\|\|z'\| \left( 1 - \kappa'(1) + \kappa'(1)u \right)$$

$$= (1 - \kappa'(1)) (\|z\|^2 + \|z'\|^2 - 2\|z\|\|z'\|) + \kappa'(1) (\|z\|^2 + \|z'\|^2 - 2\langle z, z' \rangle)$$

$$= (1 - \kappa'(1)) \|z - z'\|^2 + \kappa'(1) \|z - z'\|^2$$

$$\leq \|z - z'\|^2,$$

where we used the fact that $0 \leq \kappa'(1) \leq 1$. Note that if we make instead the assumption that $\kappa'(1) > 1$, the same derivation shows that the kernel mapping is Lipschitz with constant $\sqrt{\kappa'(1)}$. \[\square\]

We are now in shape to list three a few kernels of interest that match the above assumptions. Given two vectors $z, z' \in H_0$ with unit norm, we consider the following functions $\kappa$:

- **homogeneous Gaussian kernel**

  $$\kappa_{\text{RBF}}(\langle z, z' \rangle) = e^{\alpha(\langle z, z' \rangle - 1)} = e^{-\alpha \|z - z'\|^2} \text{ with } \alpha \leq 1.$$  

  Note that if $\alpha > 1$, the kernel mapping is expansive, but is still Lipschitz with constant $\alpha$.

- **homogeneous polynomial kernel of degree $p$\footnote{Note that the polynomial kernel of degree $p$ can be defined here: $\kappa_{\text{poly}}(\langle z, z' \rangle) = \frac{1}{p!} (p - 1 + \langle z, z' \rangle)^p$. For simplicity, we only consider the case $p = 2$ in this section.}**

  $$\kappa_{\text{poly}}(\langle z, z' \rangle) = \frac{1}{4} (1 + \langle z, z' \rangle)^2.$$  

- **homogeneous inverse polynomial kernel**

  $$\kappa_{\text{inv-poly}}(\langle z, z' \rangle) = \frac{1}{2 - \langle z, z' \rangle}.$$
• homogeneous arc-cosine kernel of degree 1 [7]:
\[ \kappa_{\text{acos}}((z, z')) = \frac{1}{\pi} (\sin(\theta) + (\pi - \theta) \cos(\theta)) \quad \text{with} \quad \theta = \arccos((z, z')). \]

• homogeneous Vovk’s real polynomial kernel of degree 3:
\[ \kappa_{\text{vovk}}((z, z')) = 1 - \frac{(z, z')^3}{3 - 3(z, z')} = \frac{1}{3} (1 + (z, z') + (z, z')^2). \]

For all of these kernels, it is indeed easy to see that the assumptions of Lemma B.1 are satisfied. We note that the inverse polynomial kernel was used in [35, 36]; below, we will use extend some of their results to characterize large subsets of functions that live in the corresponding RKHSs.

B.1 Description of the RKHSs

We consider now the kernels described previously. All of them are build with a function \( \kappa \) that admits a polynomial expansion \( \kappa(u) = \sum_{j=0}^{\infty} b_j u^j \) with \( b_j \geq 0 \) for all \( j \) and all \( u \) in \([-1, +1]\). We will now characterize these functional spaces by following the same strategy as [35, 36] for the non-homogeneous Gaussian and inverse polynomial kernels on Euclidean spaces. Using the previous Maclaurin expansion, we can define the following explicit feature map for any \( z \in \mathcal{H}_0 \):
\[
\psi(z) = \left( \sqrt{b_0} \|z\|, \sqrt{b_1} z, \sqrt{b_2} \|z\|^{-1} z \otimes z, \sqrt{b_2} \|z\|^{-1} z \otimes z, \ldots \right) = \left( \sqrt{b_j} \|z\|^{-j} z \otimes^j \right)_{j \in \mathbb{N}},
\]
where \( z \otimes^j \) denotes the tensor product of order \( j \) of the vector \( z \) in \( \mathcal{H}_0 \). Technically, the explicit mapping lives in the Hilbert space \( \oplus_{j=0}^{\infty} \mathcal{H}_0 \), where \( \oplus \) denotes the direct sum of Hilbert spaces, and with the abuse of notation that \( \otimes^0 \mathcal{H}_0 \) is simply \( \mathbb{R} \). Then, we have that \( K(z, z') = \langle \psi(z), \psi(z') \rangle \).

The next lemma extends the results of [35, 36] to our class of kernels; it shows that the RKHS contains simple “neural network” activation functions \( \sigma \) that are smooth and homogeneous.

**Lemma B.2 (Activation functions and RKHSs).** Let us consider a function \( \sigma : [-1, 1] \to \mathbb{R} \) that admits a polynomial expansion \( \sigma(u) := \sum_{j=0}^{\infty} a_j u^j \). Consider one of the previous kernels \( K \) with explicit feature map \( \psi(z) = \left( \sqrt{b_j} \|z\|^{-1-j} z \otimes^j \right)_{j \in \mathbb{N}} \), and assume that \( a_j = 0 \) if \( b_j = 0 \) for all \( j \). Define the function \( C_\sigma^2(\lambda^2) := \sum_{j=0}^{\infty} (a_j^2/b_j) \lambda^{2j} \). Let \( w \) such that \( C_\sigma^2(\|w\|^2) < \infty \). Then, the RKHS of \( K \) contains the function \( f : z \mapsto \|z\| \sigma(\langle w, z \rangle / \|z\|) \), with RKHS norm \( \|f\| \leq C_\sigma(\|w\|^2) \).

**Proof.** By first considering the restriction of \( K \) to unit-norm vectors \( z \), the proof is very similar to [35, 36]:
\[
\sigma(\langle w, z \rangle) = \sum_{j=0}^{+\infty} a_j \langle w, z \rangle^j = \sum_{j=0}^{+\infty} a_j \langle w \otimes^j, z \otimes^j \rangle = \langle \tilde{w}, \psi(z) \rangle,
\]
where
\[
\tilde{w} = \left( \frac{a_j}{\sqrt{b_j}} w \otimes^j \right)_{j \in \mathbb{N}}.
\]

Then, the norm of \( \tilde{w} \) is
\[
\|\tilde{w}\|^2 = \sum_{j=0}^{+\infty} \frac{a_j^2}{b_j} \|w \otimes^j\|^2 = \sum_{j=0}^{+\infty} \frac{a_j^2}{b_j} \|w\|^2 = C_\sigma^2(\|w\|^2) < +\infty.
\]

Using Theorem A.1, we conclude that \( f \) is in the RKHS of \( K \), with norm \( \|f\| \leq C_\sigma(\|w\|^2) \). Finally, we extend the result to non unit-norm vectors \( z \) with similar calculations and we obtain the desired result.

Some interesting activation functions are discussed in the main part of the paper, including smoothed versions of rectified linear units. The next corollary was also found to be useful in our analysis.

**Corollary B.1 (Linear functions and RKHSs).** All RKHSs considered in this section contain the linear functions of the form \( z \mapsto \langle w, z \rangle \) for all \( w \) in \( \mathcal{H}_0 \).
C.1 Proof of Lemma 1

Proof. We denote by $\bar{\Omega}$ the discrete set of sampling points considered in this lemma. The assumption on $\bar{\Omega}$ can be written as $\{u + v : u \in \bar{\Omega}, v \in S_k\} = \Omega$.

Let $B$ denote an orthonormal basis of the Hilbert space $P_k = L^2(S_k, \mathcal{H}_{k-1})$, and define the linear function $f_w = \langle w, \cdot \rangle \in \mathcal{H}_k$ for $w \in P_k$. We thus have

$$P_kx_{k-1}(u) = \sum_{w \in B} \langle w, P_kx_{k-1}(u) \rangle w$$

$$= \sum_{w \in B} f_w(P_kx_{k-1}(u))w$$

$$= \sum_{w \in B} \langle f_w, M_kP_kx_{k-1}(u) \rangle w,$$

using the reproducing property in the RKHS $\mathcal{H}_k$. Applying the pooling operator $A_k$ yields

$$A_kP_kx_{k-1}(u) = \sum_{w \in B} \langle f_w, A_kM_kP_kx_{k-1}(u) \rangle w,$$

$$= \sum_{w \in B} \langle f_w, x_k(u) \rangle w.$$

Noting that $A_kP_kx = A_k(Lv)x_{v \in S_k} = (A_kL_v)x_{v \in S_k} = (L_vA_kx)_{v \in S_k} = P_kA_kx$, with $L_vx(u) := x(u + v)$, we can evaluate at $v \in S_k$ and obtain

$$A_kx_{k-1}(u + v) = \sum_{w \in B} \langle f_w, x_k(u) \rangle w(v).$$

Thus, taking all sampling points $u \in \bar{\Omega}$ and all $v \in S_k$, we have a full view of the signal $A_kx_{k-1}$ on all of $\Omega$ by our assumption on the set $\bar{\Omega}$.

For $f \in \mathcal{H}_{k-1}$, the signal $\langle f, x_{k-1}(u) \rangle$ can then be recovered by deconvolution as follows:

$$\langle f, x_{k-1}(u) \rangle = \mathcal{F}^{-1}\left(\mathcal{F}(\langle f, A_kx_{k-1}(\cdot) \rangle) / \mathcal{F}(h_{\sigma_k})\right)(u),$$

where $\mathcal{F}$ denotes the Fourier transform. Note that the inverse Fourier transform is well-defined here because the signal $\langle f, A_kx_{k-1}(\cdot) \rangle$ is itself a convolution with $h_{\sigma_k}$, and $\mathcal{F}(h_{\sigma_k})$ is strictly positive as the Fourier transform of a Gaussian which is also a Gaussian.

By considering all elements $f$ in an orthonormal basis of $\mathcal{H}_{k-1}$, we can recover $x_{k-1}$. $x_k$ can then be reconstructed trivially by applying operators $P_k, M_k$ and $A_k$.

C.2 Proof of Lemma 2

Proof. In this proof, we drop the bar notation on all quantities for simplicity; there is indeed no ambiguity since all signals are discrete here. First, we recall that $\mathcal{H}_k$ contains all linear functions on $\mathcal{P}_k = \mathcal{H}^e_{k-1}$, thus, we may consider in particular functions $f_j, w(z) := e_k^{1/2}\langle w, z_j \rangle$ for $j \in$
\{1, \ldots, e_k\}, w \in H_{k-1}, and z = (z_1, z_2, \ldots, z_{e_k}) in P_k. Then, we may evaluate
\[
\langle f_{j, w}, s_k^{-1/2} x_k[n]\rangle = \sum_{m \in \mathbb{Z}} h_k[n s_k - m]\langle f_{j, w}, M_k P_k x_{k-1}[m]\rangle
= \sum_{m \in \mathbb{Z}} h_k[n s_k - m]\langle f_{j, w}, \varphi_k(P_k x_{k-1}[m])\rangle
= \sum_{m \in \mathbb{Z}} h_k[n s_k - m] f_{j, w}(P_k x_{k-1}[m])
= \sum_{m \in \mathbb{Z}} h_k[n s_k - m] \langle w, x_{k-1}[m + j]\rangle
= \sum_{m \in \mathbb{Z}} h_k[n s_k + j - m] \langle w, x_{k-1}[m]\rangle
= (h_k \ast \langle w, x_{k-1}\rangle)[n s_k + j],
\]
where, with an abuse of notation, \(\langle w, x_{k-1}\rangle\) is the real-valued discrete signal such that \(\langle w, x_{k-1}\rangle[n] = \langle w, x_{k-1}[n]\rangle\). Since integers of the form \((n s_k + j)\) cover all of \(\mathbb{Z}\) according to the assumption \(e_k \geq s_k\), we have a full view of the signal \((h_k \ast \langle w, x_{k-1}\rangle)\) on \(\mathbb{Z}\). We will now follow the same reasoning as in the proof of Lemma 1 to recover \(\langle w, x_{k-1}\rangle\):
\[
\langle w, x_{k-1}\rangle = \mathcal{F}^{-1} \left( \frac{\mathcal{F}(h_k \ast \langle w, x_{k-1}\rangle)}{\mathcal{F}(h_k)} \right),
\]
where \(\mathcal{F}\) is the Fourier transform. Since the signals involved there are discrete, their Fourier transform are periodic with period \(2\pi\), and we note that \(\mathcal{F}(h_k)\) is strictly positive and bounded away from zero. The signal \(x_{k-1}\) is then recovered exactly as in the proof of Lemma 1 by considering for \(w\) the elements of an orthonormal basis of the Hilbert space \(H_{k-1}\).

\(\square\)

**C.3 Proof of Proposition 3**

**Proof.** Define \((MPA)_{k,j} := M_k P_k A_{k-1} M_{k-1} P_{k-1} A_{k-2} \cdots M_j P_j A_{j-1}\). Using the fact that \(\|A_k\| \leq 1, \|P_k\| = 1\) and \(M_k\) is non-expansive, we obtain
\[
\|\Phi(L_T x) - \Phi(x)\| = \|A_n(MPA)_{n, 2} M_1 P_1 A_0 L_T x - A_n(MPA)_{n, 2} M_1 P_1 A_0 x\|
\leq \|A_n(MPA)_{n, 2} M_1 P_1 A_0 L_T x - A_n(MPA)_{n, 2} M_1 L_T P_1 A_0 x\|
+ \|A_n(MPA)_{n, 2} M_1 L_T P_1 A_0 x - A_n(MPA)_{n, 2} M_1 P_1 A_0 x\|
\leq \|(P_1 A_0, L_T)\| \|x\|
+ \|A_n(MPA)_{n, 2} M_1 L_T P_1 A_0 x - A_n(MPA)_{n, 2} M_1 P_1 A_0 x\|.
\]
Note that \(M_1\) is defined point-wise, and thus commutes with \(L_T\):
\[
M_1 L_T x(u) = \varphi_1(L_T x(u)) = \varphi_1(x(u - \tau(u))) = M_1 x(u - \tau(u)) = L_T M_1 x(u).
\]
By noticing that \(\|M_1 P_1 A_0 x\| \leq \|x\|\), we can expand the second term above in the same way. Repeating this by induction yields
\[
\|\Phi(L_T x) - \Phi(x)\| \leq \sum_{k=1}^{n} \||P_k A_{k-1}, L_T|| \|x\| + \|A_n L_T (MPA)_{n, 1} x - A_n(MPA)_{n, 1} x\|
\leq \sum_{k=1}^{n} \||P_k A_{k-1}, L_T|| \|x\| + \|A_n L_T - A_n\| \|x\|,
\]
and the result follows by decomposing \(A_n L_T = [A_n, L_T] + L_T A_n\) using the triangle’s inequality. \(\square\)

**C.4 Proof of Lemma 4**

**Proof.** The proof follows in large parts the methodology introduced by Mallat [17] in the analysis of the stability of the scattering transform. More precisely, we will follow in part the proof of Lemma
We now decompose the commutator

\[ [LcA_\sigma, L_\tau] = LcA_\sigma L_\tau - L_\tau LcA_\sigma = Lc(A_\sigma - Lc^{-1}L_\tau LcA_\sigma L_\tau^{-1})L_\tau = LcTL_\tau, \]

with \( T := A_\sigma - Lc^{-1}L_\tau LcA_\sigma L_\tau^{-1} \). Hence,

\[ ||[LcA_\sigma, L_\tau]|| \leq ||Lc||||L_\tau||||T||. \]

We have \( ||Lc|| = 1 \) since the translation operator \( Lc \) preserves the norm. Note that we have

\[ 2^{-d} \leq (1 - ||\nabla \tau||_\infty)^d \leq \det(I - \nabla \tau(u)) \leq (1 + ||\nabla \tau||_\infty)^d \leq 2^d, \]

for all \( u \in \Omega \). Thus, for \( f \in L^2(\Omega) \),

\[ ||L_\tau f||^2 = \int_\Omega |f(z - \tau(z))|^2 dz = \int_\Omega |f(u)|^2 \det(I - \nabla \tau(u))^{-1} du \]

\[ \leq (1 - ||\nabla \tau||_\infty)^{-d} ||f||^2, \]

such that \( ||L_\tau|| \leq (1 - ||\nabla \tau||_\infty)^{-d/2} \leq 2^{d/2} \). This yields

\[ ||[LcA_\sigma, L_\tau]|| \leq 2^{d/2}||T||. \]

**Kernel of \( T \).** We now show that \( T \) is an integral operator and describe its kernel. Let \( \xi = (I - \tau)^{-1} \), so that \( L_\tau^{-1}f(z) = f(\xi(z)) \) for any function \( f \) in \( L^2(\Omega) \). We have

\[ A_\sigma L_\tau^{-1}f(z) = \int h_\sigma(z - v)f(\xi(v))dv \]

\[ = \int h_\sigma(z - u + \tau(u))f(u) \det(I - \nabla \tau(u)) dv \]

using the change of variable \( v = u - \tau(u) \), giving \( \left| \frac{dv}{du} \right| = \det(I - \nabla \tau(u)) \). Then note that

\[ L_\tau^{-1}L_\tau Lc.f(z) = L_\tau Lc.f(z + c) = Lc.f(z + c - \tau(z + c)) \]

\[ = f(z - \tau(z + c) - u + \tau(u)) \det(I - \nabla \tau(u)). \]

This yields the following kernel for the operator \( T \):

\[ k(z, u) = h_\sigma(z - u) - h_\sigma(z - \tau(z + c) - u + \tau(u)) \det(I - \nabla \tau(u)). \]

A similar operator appears in Lemma E.1 of \[17\], whose kernel is identical to \[12\] when \( c = 0 \).

As in \[17\], we decompose \( T = T_1 + T_2 \), with kernels

\[ k_1(z, u) = h_\sigma(z - u) - h_\sigma((I - \nabla \tau(u))(z - u)) \det(I - \nabla \tau(u)) \]

\[ k_2(z, u) = \det(I - \nabla \tau(u))(h_\sigma((I - \nabla \tau(u))(z - u)) - h_\sigma(z - \tau(z + c) - u + \tau(u))). \]

The kernel \( k_1(z, u) \) appears in \[17\], whereas the kernel \( k_2(z, u) \) involves a shift \( c \) which is not present in \[17\]. For completeness, we include the proof of the bound for both operators, even though only dealing with \( k_2 \) requires slightly new developments.

**Bound on \( ||T_1|| \).** We can write \( k_1(z, u) = \sigma^{-d}g(u, (z - u)/\sigma) \) with

\[ g(u, v) = h(v) - h((I - \nabla \tau(u))v) \det(I - \nabla \tau(u)) \]

\[ = (1 - \det(I - \nabla \tau(u)))h((I - \nabla \tau(u))v) + h(v) - h((I - \nabla \tau(u))v). \]

\[ ^3 \text{Note that a more precise analysis may be obtained by using finer decay bounds.} \]
Using the fundamental theorem of calculus on $h$, we have
\[ h(v) - h((I - \nabla \tau(u))v) = \int_0^1 \langle \nabla h((I + (t-1)\nabla \tau(u))v), \nabla \tau(u)v \rangle dt. \]

Noticing that
\[ |(I + (t-1)\nabla \tau(u))v| \geq (1 - \|\nabla \tau\|_\infty)|v| \geq (1/2)|v|, \]
and that $\det(I - \nabla \tau(u)) \geq (1 - \|\nabla \tau\|_\infty)\alpha \geq 1 - d\|\nabla \tau\|_\infty$, we bound each term as follows
\[ \|(1 - \det(I - \nabla \tau(u)))h((I - \nabla \tau(u))v)\| \leq d\|\nabla \tau\|(1 + \frac{1}{2}|v|)^d, \]
\[ \left| \int_0^1 \langle \nabla h((I + (t-1)\nabla \tau(u))v), \nabla \tau(u)v \rangle dt \right| \leq \|\nabla \tau\|_\infty C_h |v|^d. \]

We thus have
\[ |g(u, v)| \leq \|\nabla \tau\|_\infty C_h d + C_h' |v|. \]

Using appropriate changes of variables in order to bound $\int |k_1(z, u)| du$ and $\int |k_1(z, u) dz|$, Schur’s test yields
\[ \|T_1\| \leq C_1 \|\nabla \tau\|_\infty, \quad (13) \]
with
\[ C_1 = \int_\Omega \frac{C_h d + C_h' |v|}{(1 + \frac{1}{2}|v|)^d} dv. \]

**Bound on $\|T_2\|$.** Let $\alpha(z, u) = \tau(z + c) - \tau(u) - \nabla \tau(u)(z - u)$, and note that we have
\[ |\alpha(z, u)| \leq |\tau(z + c) - \tau(u)| + |\nabla \tau(u)(z - u)| \]
\[ \leq \|\nabla \tau\|_\infty |z + c - u| + \|\nabla \tau\|_\infty |z - u| \]
\[ \leq \|\nabla \tau\|_\infty (|c| + 2|z - u|). \quad (14) \]

The fundamental theorem of calculus yields
\[ k_2(z, u) = -\det(I - \nabla \tau(u)) \int_0^1 \langle \nabla h_\sigma(z - \tau(z + c) - u + \tau(u) - t\alpha(z, u)), \alpha(z, u) \rangle dt. \]

We note that $|\det(I - \nabla \tau(u))| \leq 2^d$, and $\nabla h_\sigma(v) = \sigma^{-d-1} \nabla h(v/\sigma)$. Using the change of variable $z' = (z - u)/\sigma$, we obtain
\[ \int |k_2(z, u)| dz \leq \]
\[ 2^d \int_0^1 \left| \frac{\nabla h\left( z' + \frac{\tau(u + \sigma z' + c) - \tau(u) - t\alpha(z, u)}{\sigma} \right)}{\sigma} \right| \langle \frac{\alpha(u + \sigma z', u)}{\sigma} \rangle \ dt dz'. \]

We can use the upper bound \((14)\):
\[ \left| \frac{\alpha(u + \sigma z', u)}{\sigma} \right| \leq \|\nabla \tau\|_\infty (\kappa + 2|z'|). \quad (15) \]

Separately, we have $|\nabla h(v(z'))| \leq C'_h/(1 + |v(z')|)^d$, with
\[ v(z') := z' + \frac{\tau(u + \sigma z' + c) - \tau(u) - t\alpha(u + \sigma z', u)}{\sigma}. \]

For $|z'| > 2\kappa$, we have
\[ \left| \frac{\tau(u + \sigma z' + c) - \tau(u) - t\alpha(u + \sigma z', u)}{\sigma} \right| = \left| t\nabla \tau(u) z' + (1 - t) \frac{\tau(u + \sigma z' + c) - \tau(u)}{\sigma} \right| \]
\[ \leq t\|\nabla \tau\|_\infty |z'| + (1 - t)\|\nabla \tau\|_\infty (|z'| + \kappa) \leq \frac{3}{2}\|\nabla \tau\|_\infty |z'| \leq \frac{3}{4}|z'|, \]

and hence, using the reverse triangle inequality, \( |v(z')| \geq |z'|-\frac{3}{4}|z'| = \frac{1}{4}|z'| \). This yields the upper bound

\[
|\nabla h(v(z'))| \leq \begin{cases} 
  C_h', & \text{if } |z'| \leq 2\kappa \\
  \frac{C_h'}{(1+\frac{1}{4}|z'|)^{d+2}}, & \text{if } |z'| > 2\kappa.
\end{cases}
\]  

(16)

Combining these two bounds, we obtain

\[
\int |k_2(z,u)|dz \leq C_2\|\nabla \tau\|_{\infty},
\]

with

\[
C_2 := 2^d C_h' \left( \int_{|z'|<2\kappa} (\kappa + 2|z'|)dz' + \int_{|z'|>2\kappa} \frac{\kappa + 2|z'|}{(1+\frac{1}{4}|z'|)^{d+2}}dz' \right).
\]

Note that the dependence of the first integral on \( \kappa \) is of order \( k^{d+1} \). Following the same steps with the change of variable \( u' = (z-u)/\sigma \), we obtain the bound \( \int |k_2(z,u)|du \leq C_2\|\nabla \tau\|_{\infty} \). Schur’s test then yields

\[
\|T_2\| \leq C_2\|\nabla \tau\|_{\infty}.
\]  

(17)

We have thus proven

\[
\|[-LcA_\sigma, L_\tau]\| \leq 2^{d/2}\|T\| \leq 2^{d/2}(C_1 + C_2)\|\nabla \tau\|_{\infty}.
\]

\[\Box\]

### D Proof of Proposition 7 and Construction of CNNs in the RKHS

In this section, we describe the space of functions (RKHS) \( \mathcal{H}_{K_n} \) associated to the kernel \( K_n(x,x') = \langle \Phi_n(x), \Phi_n(x') \rangle \) where \( \Phi_n \) is the final representation of Eq. (3), in particular showing it contains the set of CNNs with activations described in Section B.7.

**Construction of a CNN in the RKHS.** Let us consider the definition of the CNN presented in Section 5. We will show that it can be seen as a point in the RKHS of \( K_n \). According to Lemma B.2, we consider \( \mathcal{H}_k \) that contains all functions of the form \( z \in \mathcal{P}_k \mapsto \|z\|\sigma((w,z)/\|z\|), w \in \mathcal{P}_k \).

Then, we define the initial quantities \( f_1^i \in \mathcal{H}_1, g_1^i \in \mathcal{P}_1 \) for \( i = 1, \ldots, p_1 \) such that

\[
f_1^i(z) = \|z\|\sigma(g_0^i, z)/\|z\| \quad \text{for } z \in \mathcal{P}_1,
\]

and we recursively define, from layer \( k-1 \), the quantities \( f_k^i \in \mathcal{H}_k, g_k^i \in \mathcal{P}_k \) for \( i = 1, \ldots, p_k \):

\[
g_k^i(v) = \sum_{j=1}^{p_{k-1}} w_{k,j}^i(v) f_{k-1}^j \quad \text{where} \quad w_{k,j}^i(v) = (w_{k,j}^i(v))_{j=1,\ldots,p_{k-1}}
\]

\[
f_k^i(z) = \|z\|\sigma(g_k^i, z)/\|z\| \quad \text{for } z \in \mathcal{P}_k.
\]

Then, we will show that \( \hat{z}_k^i(u) = f_k^i(P_k x_{k-1}(u)) = \langle f_k^i, M_k P_k x_{k-1}(u) \rangle \) which correspond to feature maps at layer \( k \) and index \( i \) in a CNN. Indeed, this is easy to see for \( k = 1 \) by construction with
filters \( w^i_k(v) \), and for \( k \geq 2 \), we have

\[
\tilde{z}^i_k(u) = n_k(u) \sigma \left( \langle w^i_k, P_kz_{k-1}(u) \rangle / n_k(u) \right) = n_k(u) \sigma \left( \langle w^i_k, P_kA_k \tilde{z}^i_{k-1}(u) \rangle / n_k(u) \right)
\]

\[
= n_k(u) \sigma \left( \frac{1}{n_k(u)} \sum_{j=1}^{p_k} \int_{S_k} w^{ij}_k(v) A_k \tilde{z}^i_{k-1}(u + v) dv_k(u) \right)
\]

\[
= n_k(u) \sigma \left( \frac{1}{n_k(u)} \sum_{j=1}^{p_k} \int_{S_k} w^{ij}_k(v) \langle f^j_{k-1}, A_kM_kP_kx_{k-2}(u + v) \rangle dv_k(v) \right)
\]

\[
= n_k(u) \sigma \left( \frac{1}{n_k(u)} \int_{S_k} \langle g^j_k(v), A_kM_kP_kx_{k-2}(u + v) \rangle dv_k(v) \right)
\]

\[
= n_k(u) \sigma \left( \frac{1}{n_k(u)} \int_{S_k} \langle g^j_k(v), x_{k-1}(u + v) \rangle dv_k(v) \right)
\]

\[
= n_k(u) \sigma \left( \frac{1}{n_k(u)} \langle g^j_k(v), P_kx_{k-1}(u) \rangle \right)
\]

\[
= f^j_k(P_kx_{k-1}(u)),
\]

where \( n_k(u) := \| P_kx_{k-1}(u) \| \). Note that we have used many times the fact that \( A_k \) operates on each channel independently when applied to a finite-dimensional map.

The final prediction function is of the form \( f_\sigma(x_0) = \langle w_{n+1}, z_n \rangle \) with \( w_{n+1} \) in \( L^2(\Omega, \mathbb{R}^p) \). Then, we can define the following function \( g_\sigma \) in \( L^2(\Omega, \mathcal{H}_n) \) such that

\[
g_\sigma(u) = \sum_{j=1}^{p_n} w^{ij}_{n+1}(u) f^j_n,
\]

which yields

\[
\langle g_\sigma, x_n \rangle = \sum_{j=1}^{p_n} \int_{\Omega} w^{ij}_{n+1}(u) \langle f^j_n, x_n(u) \rangle du
\]

\[
= \sum_{j=1}^{p_n} \int_{\Omega} w^{ij}_{n+1}(u) \langle f^j_n, A_nM_nP_nx_{n-1}(u) \rangle du
\]

\[
= \sum_{j=1}^{p_n} \int_{\Omega} w^{ij}_{n+1}(u) A_n \tilde{z}^i_n(u) du
\]

\[
= \sum_{j=1}^{p_n} \int_{\Omega} w^{ij}_{n+1}(u) z^j_n(u) du
\]

\[
= \sum_{j=1}^{p_n} \langle w^{ij}_{n+1}, z^j_n \rangle = f_\sigma(x_0),
\]

which corresponds to a linear layer after pooling. Since the RKHS of \( \mathcal{K}_n \) contains all functions of the form \( f(x_0) = \langle g, \Phi_n(x_0) \rangle \) = \( \langle g, x_n \rangle \), for \( g \) in \( L^2(\Omega, \mathcal{H}_n) \), we have that \( f_\sigma \) is in the RKHS.

**Norm of the CNN.** As shown in Section B.1, the RKHS norm of a function \( f : z \in \mathcal{P}_k \mapsto \| z \| \sigma(\| w, z \|/\| z \|) \) in \( \mathcal{H}_k \) is bounded by \( C_\sigma(\| w \|^2) \), where \( C_\sigma \) depends on the activation \( \sigma \). We then
have
\[ \|f^i_1\|_2^2 \leq C_2^2(\|w^i_1\|_2^2) \quad \text{where} \quad \|w^i_1\|_2^2 = \int_{S_1} \|w^i_1(v)\|d\nu_1(v) \]
\[ \|f^i_k\|_2^2 \leq C_2^2(\|g^i_k\|_2^2) \]
\[ \|g^i_k\|_2^2 = \int_{S_k} \|\sum_{j=1}^{p_{k-1}} w^{ij}_k(v) f^j_{k-1}\|^2 d\nu_k(v) \]
\[ \leq p_{k-1} \sum_{j=1}^{p_{k-1}} \left( \int_{S_k} |w^{ij}_k(v)|^2 d\nu_k(v) \right) \|f^j_{k-1}\|_2^2 \]
\[ = p_{k-1} \sum_{j=1}^{p_{k-1}} \|w^{ij}_k\|_2^2 \|f^j_{k-1}\|_2^2, \]
where in the last inequality we use \( \|a_1 + \ldots + a_n\|^2 \leq n(\|a_1\|^2 + \ldots + \|a_n\|^2) \). Since \( C_2^2 \) is monotonically increasing (typically exponentially in its argument), we have for \( k = 1, \ldots, n - 1 \) the recursive relation
\[ \|f^i_k\|_2^2 \leq C_2^2 \left( p_{k-1} \sum_{j=1}^{p_{k-1}} \|w^{ij}_k\|_2^2 \|f^j_{k-1}\|_2^2 \right). \]
The norm of the final prediction function \( f \in L^2(\Omega, \mathcal{H}_n) \) is bounded as follows, using similar arguments as well as Theorem A.1,
\[ \|f_\sigma\|_2 \leq \|w_{n+1}\|_2 \leq p_n \sum_{j=1}^{p_n} \left( \int_{\Omega} |w^i_{n+1}(u)|^2 du \right) \|f^i_n\|_2^2. \]