Learning with convolution and pooling operations in kernel methods

Theodor Misiakiewicz
Department of Statistics
Stanford University
Stanford, CA 94305
misiakie@stanford.edu

Song Mei
Department of Statistics
University of California Berkeley
Berkeley, CA 94720
songmei@berkeley.edu

Abstract

Recent empirical work has shown that hierarchical convolutional kernels inspired by convolutional neural networks (CNNs) significantly improve the performance of kernel methods in image classification tasks. A widely accepted explanation for their success is that these architectures encode hypothesis classes that are suitable for natural images. However, understanding the precise interplay between approximation and generalization in convolutional architectures remains a challenge. In this paper, we consider the stylized setting of covariates (image pixels) uniformly distributed on the hypercube, and characterize exactly the RKHS of kernels composed of single layers of convolution, pooling, and downsampling operations. We use this characterization to compute sharp asymptotics of the generalization error for any given function in high-dimension. In particular, we quantify the gain in sample complexity brought by enforcing locality with the convolution operation and approximate translation invariance with average pooling. Notably, these results provide a precise description of how convolution and pooling operations trade off approximation with generalization power in one layer convolutional kernels.

1 Introduction

Convolutional neural networks (CNNs) have become essential elements of the deep learning toolbox, achieving state-of-the-art performance in many computer vision tasks [27, 30]. CNNs are constructed by stacking convolution and pooling layers, which were shown to be paramount to their empirical success [31]. A widely accepted hypothesis to explain their favorable properties is that these architectures successfully encode useful properties of natural images: locality and compositionality of the data, stability by local deformations, and translation invariance. While some theoretical progress has been made in studying the approximation and generalization benefits brought by convolution and pooling operations [6, 15, 16], our mathematical understanding of the interaction between network architecture, image distribution, and efficient learning remains limited.

Consider $\boldsymbol{x} \in \mathbb{R}^d$ an input signal, which we can think of as a grayscale pixel representation of an image. For mathematical convenience, we will consider one-dimensional images with cyclic convention $x_{d+i} := x_i$, and denote $\boldsymbol{x}_{(k)} = (x_k, x_{k+1}, \dots, x_{k+q-1})$ the k-th patch of the signal $\boldsymbol{x}$, $k \in [d]$, with patch size $q \leq d$. Most of our results can be extended to two-dimensional images.

We further consider a simple convolutional neural network composed of a single convolution layer followed by local average pooling and downsampling. The network first computes the nonlinear convolution of N filters $\boldsymbol{w}_1, \dots, \boldsymbol{w}_N \in \mathbb{R}^q$ with the image patches $\boldsymbol{x}_{(k)}$. The outputs of the convolution operation $\sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}_{(k)} \rangle)$ are then averaged locally over segments of length ω (local average pooling). This pooling operation is followed by downsampling which extracts one out of every Δ output coordinates (for simplicity, Δ is assumed to be a divisor of d). Finally, the results are

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

combined linearly using coefficients $(a_{ik})_{i\in[N],k\in[d/\Delta]}$:

$$f_{\mathsf{CNN}}(\boldsymbol{x};\boldsymbol{a},\boldsymbol{\Theta}) = \sqrt{\frac{\Delta}{N\omega d}} \sum_{i\in[N]} \sum_{k\in[d/\Delta]} a_{ik} \sum_{s\in[\omega]} \sigma\left(\langle \boldsymbol{w}_i, \boldsymbol{x}_{(k\Delta+s)}\rangle\right) . \tag{CNN-AP-DS}$$

Note that pooling and downsampling operations are often tied together in the literature. However in this work we will treat these two operations separately.

In the formula above, different values for q, ω, Δ lead to different architectures with vastly different behaviors. For example, when $q = \Delta = d$ and $\omega = 1$, we recover a two-layer *fully-connected* neural network $f_{\mathsf{FC}}(\boldsymbol{x};\boldsymbol{a},\boldsymbol{\Theta}) = N^{-1/2}\sum_{i\in[N]} a_i\sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle)$ which has the universal approximation property at large N. When $\omega = \Delta = 1$ and $q < d$, the network is *locally connected* $f_{\mathsf{LC}}(\boldsymbol{x};\boldsymbol{a},\boldsymbol{\Theta}) = N^{-1/2}\sum_{i\in[N],k\in[d]} a_{ik}\sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}_{(k)}\rangle)$, and not a universal approximator anymore: however, f_{LC} vastly outperforms f_{FC} in some cases [33]. For $\omega > 1$, local pooling enables learning functions that are locally invariant by translations more efficiently than without pooling. For $\omega = d$ (*global pooling*), the network only fits functions fully invariant by cyclic translations.

The aim of this paper is to formalize and quantify the *interplay between the target function class and the statistical efficiency brought by these different architectures*. As a concrete first step in this direction, we consider kernel models that are naturally associated to (CNN-AP-DS) through the neural tangent kernel perspective [18, 28]. Kernel methods have the advantage of 1) being tractable—leaving the computational issue of learning CNNs aside; 2) having well-understood approximation and generalization properties, which depends on the eigendecomposition of the kernel and the alignment between the target function and the RKHS [12, 47] (see Appendices B and C for background). While kernel models only describe neural networks in the lazy training regime [1, 13, 19, 20, 49] and miss important properties of deep learning, such as feature learning, architecture choice already plays a crucial role to learn efficiently 'image-like' functions in the fixed-feature regime.

Neural tangent kernels are obtained by linearizing the associated neural networks. Here we consider the tangent kernel associated to the network f_{CNN} (c.f. Appendix A.2 for a detailed derivation):

$$H^{\mathsf{CK}}_{\omega,\Delta}(\boldsymbol{x},\boldsymbol{y}) = \frac{\Delta}{d\omega} \sum_{k\in[d/\Delta]} \sum_{s,s'\in[\omega]} h\big(\langle \boldsymbol{x}_{(k\Delta+s)}, \boldsymbol{y}_{(k\Delta+s')}\rangle/q\big) , \tag{CK-AP-DS}$$

where $h : \mathbb{R} \to \mathbb{R}$ is related to the activation function σ in (CNN-AP-DS). As a linearization of CNNs, the kernel (CK-AP-DS) inherits some of the favorable properties of convolution, pooling, and downsampling operations. Indeed, a line of work [2, 32, 35, 36, 45] showed that, though performing slightly worse than CNNs, such (hierarchical) convolutional kernels have empirically outperformed the former state-of-the-art kernels. For instance, these kernels achieved test accuracy around $87\% - 90\%$ on CIFAR-10 (the state-of-the-art CNNs can achieve test accuracy 99%), against 79.6% for the best former unsupervised feature-extraction method [14].

In this paper, we will further consider a stylized setting with input signal distribution $\boldsymbol{x} \sim \mathrm{Unif}(\mathscr{Q}^d)$ (uniform distribution over $\mathscr{Q}^d := \{-1,+1\}^d$ the discrete hypercube in d dimensions). This simple choice allows for a complete characterization of the eigendecomposition of $H^{\mathsf{CK}}_{\omega,\Delta}$, thanks to all patches having same marginal distribution $\boldsymbol{x}_{(k)} \sim \mathrm{Unif}(\mathscr{Q}^q)$. We will be particularly interested in four specific choices of (q,ω,Δ) in (CK-AP-DS):

$$H^{\mathsf{FC}}(\boldsymbol{x},\boldsymbol{y}) = h\big(\langle \boldsymbol{x},\boldsymbol{y}\rangle/d\big) , \tag{FC}$$

$$H^{\mathsf{CK}}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{d}\sum_{k\in[d]} h\big(\langle \boldsymbol{x}_{(k)},\boldsymbol{y}_{(k)}\rangle/q\big) , \tag{CK}$$

$$H^{\mathsf{CK}}_{\omega}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{d\omega}\sum_{k\in[d]} \sum_{s,s'\in[\omega]} h\big(\langle \boldsymbol{x}_{(k+s)},\boldsymbol{y}_{(k+s')}\rangle/q\big) , \tag{CK-AP}$$

$$H^{\mathsf{CK}}_{\mathsf{GP}}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{d}\sum_{k,k'\in[d]} h\big(\langle \boldsymbol{x}_{(k)},\boldsymbol{y}_{(k')}\rangle/q\big) . \tag{CK-GP}$$

These kernels are respectively the neural tangent kernels of a fully-connected network f_{FC} (FC), a convolutional network f_{LC} (CK), a convolutional network followed by local average pooling (CK-AP)

and a convolutional network followed by global pooling (CK-GP). We will further be interested in (CK-GP) with patch size $q = d$, which we denote $H^{\mathsf{FC}}_{\mathsf{GP}}$: this corresponds to a convolutional kernel with full-size patches $q = d$, followed by global pooling.

In this paper, we first characterize the reproducing kernel Hilbert space (RKHS) of these convolutional kernels, and then investigate their generalization properties in the regression setup. More specifically, assume $\{(\boldsymbol{x}_i, y_i)\}_{i\le n}$ are n i.i.d. samples with $\boldsymbol{x}_i \sim \mathrm{Unif}(\mathscr{Q}^d)$ and $y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i$. Here $f_\star \in L^2(\mathscr{Q}^d)$ and $(\varepsilon_i)_{i\le n}$ are independent errors with mean zero and variance bounded by σ_ε^2. We will focus on the generalization error of kernel ridge regression (KRR) (see Appendix B.1 for general kernel methods). In particular, given a kernel function $H : \mathscr{Q}^d \times \mathscr{Q}^d \to \mathbb{R}$ and a regularization parameter $\lambda \ge 0$, the KRR estimator is the solution of the tractable convex problem

$$\hat{f}_\lambda = \arg\min_{f\in\mathcal{H}} \Big\{ \sum_{i\in[n]} \big(y_i - f(\boldsymbol{x}_i)\big)^2 + \lambda \|f\|_{\mathcal{H}}^2 \Big\}, \tag{KRR}$$

where $\mathcal{H}$ is the RKHS associated to H with RKHS norm $\|\cdot\|_{\mathcal{H}}$. We denote the test error with square loss by $R(f_\star, \hat{f}_\lambda) = \mathbb{E}_{\boldsymbol{x}}\{(f_\star(\boldsymbol{x}) - \hat{f}_\lambda(\boldsymbol{x}))^2\}$. We will sometimes consider the expected test error $\mathbb{E}_{\boldsymbol{\varepsilon}}\{R(f_\star, \hat{f}_\lambda)\}$, where expectation is taken with respect to noise $\boldsymbol{\varepsilon} = (\varepsilon_i)_{i\le n}$ in the training data.

The generalization properties of kernels H^{FC} and $H^{\mathsf{FC}}_{\mathsf{GP}}$ were recently studied in [7, 39]. In particular, they showed that global pooling (kernel $H^{\mathsf{FC}}_{\mathsf{GP}}$) leads to a gain of a factor d in sample complexity when fitting cyclic invariant functions, but still suffers from the curse of dimensionality ($H^{\mathsf{FC}}_{\mathsf{GP}}$ only fits very smooth functions in high-dimension). More precisely, Mei et al. [39] considered the high-dimensional framework of [38] and showed the following: KRR with H^{FC} requires $n \approx d^\ell$ samples to fit degree-ℓ cyclic polynomials, while KRR with $H^{\mathsf{FC}}_{\mathsf{GP}}$ only needs $n \approx d^{\ell-1}$. To enable milder dependence on the dimension d, further structural assumptions on the kernel and the target function should be considered (for instance, in this paper, we use the kernel H^{CK} and consider 'local' functions).

1.1 Summary of main results

Our contributions are two-fold. First, we describe the RKHS associated with the convolutional kernel (CK-AP-DS) in the stylized setting $\boldsymbol{x} \sim \mathrm{Unif}(\mathscr{Q}^d)$, which provides a fully explicit picture of the roles of convolution, pooling and downsampling operations in approximating specific classes of functions. Second, we provide sharp asymptotics for the generalization error of KRR in high-dimension, given any target function and one of the kernels described in the introduction[1]. These asymptotics are obtained rigorously using the framework of [38] (see Appendix C for background). For completeness, we also include bounds on the KRR test error in the classical fixed-dimension setting with capacity/source assumptions (see Appendix C for limitations of this classical approach).

We summarize our results below. Define the q-local function class $L^2(\mathscr{Q}^d, \mathrm{Loc}_q)$ and the cyclic q-local function class $L^2(\mathscr{Q}^d, \mathrm{CycLoc}_q)$ (subspace of $L^2(\mathscr{Q}^d, \mathrm{Loc}_q)$ consisting of cyclic-invariant functions) as follows:

$$L^2(\mathscr{Q}^d, \mathrm{Loc}_q) = \Big\{ f \in L^2(\mathscr{Q}^d) : \exists \{g_k\}_{k\in[d]} \subseteq L^2(\mathscr{Q}^q), f(\boldsymbol{x}) = \sum_{k\in[d]} g_k(\boldsymbol{x}_{(k)}) \Big\}, \tag{LOC}$$

$$L^2(\mathscr{Q}^d, \mathrm{CycLoc}_q) = \Big\{ f \in L^2(\mathscr{Q}^d) : \exists g \in L^2(\mathscr{Q}^q), f(\boldsymbol{x}) = \sum_{k\in[d]} g(\boldsymbol{x}_{(k)}) \Big\}. \tag{CYC-LOC}$$

One-layer convolutional layer. The RKHS of H^{CK} is equal to $L^2(\mathscr{Q}^d, \mathrm{Loc}_q)$: kernel methods with H^{CK} can only fit the projection $\mathsf{P}_{\mathrm{Loc}_q} f_*$ of the target function onto $L^2(\mathscr{Q}^d, \mathrm{Loc}_q)$. For a sample size $n \asymp dq^{\ell-1}$, KRR fits exactly a degree-ℓ polynomial approximation to $\mathsf{P}_{\mathrm{Loc}_q} f_*$. In particular, for $q \ll d$, the convolution kernel H^{CK} is much more sample efficient than the standard inner-product kernel H^{FC} for fitting functions in $L^2(\mathscr{Q}^d, \mathrm{Loc}_q)$ (sample sizes $dq^{\ell-1} \ll d^\ell$ for fitting a degree-ℓ polynomials). *The convolution operation breaks the curse of dimensionality by restricting the RKHS to local functions.*

[1] Note that we modify slightly H^{CK}_ω to simplify the derivation of the high-dimension asymptotics. However, we believe such a simplification to be unecessary. The fixed-dimension bounds do not require such a simplification.

To fit a degree ℓ polynomial	H^{FC}	$H^{\mathsf{FC}}_{\mathsf{GP}}$	H^{CK}	H^{CK}_{ω}	$H^{\mathsf{CK}}_{\mathsf{GP}}$
Sample complexity	d^ℓ	$d^{\ell-1}$	$dq^{\ell-1}$	$dq^{\ell-1}/\omega$	$q^{\ell-1}$

Table 1: Sample size n required to fit a q-local cyclic-invariant polynomial of degree ℓ using kernel ridge regression (KRR) with the 5 different kernels of interest in this paper.

Average pooling. The RKHS of H^{CK}_{ω} is still constituted of q-local functions $f_* \in L^2(\mathscr{Q}^d, \mathrm{Loc}_q)$, but penalizes differently the frequency components $f_{*,j}(\boldsymbol{x})$ by reweighting their eigenspaces by a factor κ_j, where $f_{*,j}(\boldsymbol{x}) = \sum_{k\in[d]} \rho_j^k f_*(t_k \cdot \boldsymbol{x})$ with $\rho_j = e^{\frac{2i\pi j}{d}}$ and $t_k \cdot \boldsymbol{x} = (x_{k+1}, \dots, x_d, x_1, \dots, x_k)$ denotes the k-shift. As ω increases, local pooling penalizes more and more heavily the high-frequency components ($\kappa_j \ll 1$), while making low-frequency components statistically easier to learn ($\kappa_j \gg 1$). For global pooling $\omega = d$, $H^{\mathsf{CK}}_{\mathsf{GP}}$ only learns cyclic local functions $L^2(\mathscr{Q}^d, \mathrm{CycLoc}_q)$ and enjoy a factor d gain in statistical complexity compared to H^{CK} (sample sizes $q^{\ell-1} \ll dq^{\ell-1}$ to learn a degree-ℓ polynomial). *Local pooling biases learning towards functions that are stable by small translations.*

Downsampling. When $\Delta \leq \omega$, downsampling after average pooling leaves the low-frequency eigenspaces of H^{CK}_{ω} stable. In particular, the downsampling operation does not modify the statistical complexity of learning low-frequency functions in one-layer kernels, while being potentially beneficial in further layers in deep convolutional kernels.

These theoretical results answer the following question: *given a target function and a sample size n, what is the impact of the architecture on the test error?* For example, Table 1 shows how the architecture modify the sample size required to achieve small test error when learning a degree-ℓ polynomial in $L^2(\mathscr{Q}^d, \mathrm{CycLoc}_q)$.

There are two important model assumptions in this paper, which deserve some discussions:

One-layer convolutional kernel (CK): extra layers allow for hierarchical interactions between the patches (see for example [6]). However, we believe that the main insights on the approximation and statistical trade-off are already captured in the one-layer case (see [48] for multi-layer but independent patches). Note that depth might be less important for CKs than for CNNs: the one-layer CK considered in this paper achieves 80.9% accuracy on CIFAR-10 [6] (versus 79.6% in [14]) and 3-layers CK achieves 88.2% accuracy [6] (versus 90% for the best multi-layer CK [45]). See Appendix A.5 for a discussion on how our results could be extended to 2-layers.

Data uniform on the hypercube: this choice is motivated by our goal of deriving rigorous fine-grained approximation and generalization errors, which requires to diagonalize the kernel (CK-AP-DS). More general data distributions either require strong assumptions (independent patches [43, 48]), loose minmax bounds on the generalization error (e.g., classical source/capacity assumptions) or non-rigorous statistical physics heuristics [22].

The rest of the paper is organized as follows. We discuss related work in Section 1.2. In Section 2, we present our main results on convolutional kernels and describe precisely the roles of convolution, pooling and downsampling operations. Finally, we present a numerical simulation on synthetic data in Section 3 and conclude in Section 4. Some details and discussions are deferred to Appendix A.

1.2 Related work

Convolutional kernels have been considered in [6, 32, 35, 36, 45, 46]. In particular, they showed that these architectures achieve good results in image classification (90% accuracy on Cifar10) and that pooling and downsampling were necessary for their good performance [32].

The generalization error of kernel ridge regression (KRR) has been well-studied in both the fixed dimension regime [47, Chap. 13], [12] and the high-dimensional regimes [21, 23, 24, 34, 39, 48]. These results show that the generalization error depends on the eigenvalues and eigenfunctions of the kernel, and the alignment of the kernel with the target function.

Recently, a few theoretical work have considered the generalization properties of invariant kernels and convolutional kernels [7, 22, 39, 43, 48]. In particular, a concurrent work [48] considers sharp asymptotics of the KRR test error using the framework of [38] for certain hierarchical convolutional

kernels under the strong assumption of non-overlapping patches (whereas we consider the more natural architecture of overlapping patches). They arrive at a similar trade-off between approximation and generalization power in convolutional kernels, which they call 'eigenspace restructuring principle': given a finite statistical budget (i.e., a sample size n), convolutional architectures allocate the 'eigenvalue mass' by weighting differently the eigenspaces. Favero et al. [22] considers a one-layer convolutional kernel with and without global pooling and derive asymptotic rates in sample size n in a student-teacher scenario using statistical physics heuristics and a Gaussian equivalence conjecture. In particular, they show that locality rather than translation-invariance breaks the curse of dimensionality. Our goal in this paper is different: we derive *rigorous quantitative bounds* that give separation in generalization power between different architectures.

See [33, 37] for more theoretical results on the separation between convolutional and fully connected neural networks, and [10, 16] for the inductive bias of pooling operations in convolutional neural networks.

2 Main results

We start by introducing some background on functions on the hypercube and eigendecomposition of kernel operators in Section 2.1. We first consider a kernel with a single convolution layer in Section 2.2, and characterize its eigendecomposition and generalization properties. We then show how these results are modified when applying local average pooling and downsampling in Section 2.3.

2.1 Functions on the hypercube and eigendecomposition of kernel operators

Recall that we work on the d-dimensional hypercube $\mathscr{Q}^d := \{-1,+1\}^d$. Let $L^2(\mathscr{Q}^d) = L^2(\mathscr{Q}^d, \mathrm{Unif})$ be the 2^d-dimensional vector space of all functions $f : \mathscr{Q}^d \to \mathbb{R}$, with scalar product $\langle f, g\rangle_{L^2} := \mathbb{E}_{\boldsymbol{x}\sim\mathrm{Unif}(\mathscr{Q}^d)}[f(\boldsymbol{x})g(\boldsymbol{x})]$. Let $\|\cdot\|_{L^2}$ be the norm associated with the scaler product. We introduce the set of Fourier functions $\{Y_S^{(d)}(\boldsymbol{x})\}_{S\subseteq[d]}$ which forms an orthonormal basis of $L^2(\mathscr{Q}^d)$. For any subset $S \subseteq [d]$, the Fourier function is defined as $Y_S^{(d)}(\boldsymbol{x}) := \prod_{i\in S} x_i$ with the convention that $Y_\emptyset^{(d)} := 1$ (it is easy to verify that $\langle Y_S^{(d)}, Y_{S'}^{(d)}\rangle_{L^2} = \mathbf{1}_{S=S'}$). We will omit the superscript (d) which will be clear from context and write $Y_S := Y_S^{(d)}$.

Consider a nonnegative definite kernel function $H : \mathscr{Q}^p \times \mathscr{Q}^p \to \mathbb{R}$ ($p = d$ or q in this paper) with associated integral operator $\mathbb{H} : L^2(\mathscr{Q}^p) \to L^2(\mathscr{Q}^p)$ defined as $\mathbb{H}f(\boldsymbol{u}) = \mathbb{E}_{\boldsymbol{v}}\{h(\boldsymbol{u},\boldsymbol{v})f(\boldsymbol{v})\}$ with $\boldsymbol{v} \sim \mathrm{Unif}(\mathscr{Q}^p)$. By spectral theorem of compact operators, there exists an orthonormal basis $\{\psi_j\}_{j\ge1}$ of $L^2(\mathscr{Q}^p)$ and nonnegative eigenvalues $(\lambda_j)_{j\ge1}$ such that $\mathbb{H} = \sum_{j\ge1}\lambda_j\psi_j\psi_j^*$ (i.e., $H(\boldsymbol{u},\boldsymbol{v}) = \sum_{j\ge1}\lambda_j\psi_j(\boldsymbol{u})\psi_j(\boldsymbol{v})$ for any $\boldsymbol{u},\boldsymbol{v} \in L^2(\mathscr{Q}^p)$).

The most widespread example are *inner-product* kernels defined as $H(\boldsymbol{u},\boldsymbol{v}) := h(\langle\boldsymbol{u},\boldsymbol{v}\rangle/p)$ for some function $h : \mathbb{R} \to \mathbb{R}$. Inner-product kernels have the following simple eigendecomposition in $L^2(\mathscr{Q}^p)$ (taking here $\boldsymbol{u},\boldsymbol{v} \in \mathscr{Q}^p$):

$$h\big(\langle\boldsymbol{u},\boldsymbol{v}\rangle/p\big) = \sum_{\ell=0}^{p}\xi_{p,\ell}(h)\sum_{S\subseteq[p],|S|=\ell}Y_S(\boldsymbol{u})Y_S(\boldsymbol{v}), \tag{1}$$

where $\xi_{p,\ell}(h)$ is the ℓ-th Gegenbauer coefficient of $h(\cdot/\sqrt{p})$ in dimension p, i.e.,

$$\xi_{p,\ell}(h) = \mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(\mathscr{Q}^p)}\big[h(\langle\boldsymbol{u},\boldsymbol{e}\rangle/p)Q_\ell^{(p)}(\langle\boldsymbol{u},\boldsymbol{e}\rangle)\big], \tag{2}$$

for $\boldsymbol{e} \in \mathscr{Q}^p$ arbitrary and $Q_\ell^{(p)}$ the degree-ℓ Gegenbauer polynomial on $\mathscr{Q}^p$ (see Appendix D for details). Note that $(\xi_{p,\ell})_{0\le\ell\le q}$ are non-negative by positive semidefiniteness of the kernel. We will write $\xi_{p,\ell} := \xi_{p,\ell}(h)$ and use extensively the decomposition identity (1) in the rest of the paper.

2.2 One-layer convolutional kernel

We first consider the convolutional kernel H^{CK} (CK) given by a one-layer convolution layer with patch size q and inner-product kernel function $h : \mathbb{R} \to \mathbb{R}$:

$$H^{\mathsf{CK}}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{d}\sum_{k=1}^{d} h\left(\langle\boldsymbol{x}_{(k)},\boldsymbol{y}_{(k)}\rangle/q\right), \tag{3}$$

where we recall that $\boldsymbol{x}_{(k)} = (x_k, \dots, x_{k+q-1}) \in \mathscr{Q}^q$ is the k'th patch of the image with size q.

Before stating the eigendecomposition of H^{CK}, we introduce some notations. For any subset $S \subseteq [d]$, denote $\gamma(S)$ the diameter of S with cyclic convention, i.e., $\gamma(S) = \max\{\min\{\text{mod}(j-i,d)+1, \text{mod}(i-j,d)+1\} : i,j \in S\}$ (e.g., $\gamma(\{2,d\}) = 3$). For any integer $\ell \leq q$, consider the set $\mathcal{E}_\ell = \{S \subseteq [d] : |S| = \ell, \gamma(S) \leq q\}$ of all subsets of $[d]$ of size ℓ with diameter less or equal to q. We will assume throughout this paper that $q \leq d/2$ to avoid additional overlap between sets.

Proposition 1 (Eigendecomposition of H^{CK})**.** *Let H^{CK} be a convolutional kernel as defined in Eq.* (3)*. Then H^{CK} admits the following eigendecomposition:*

$$H^{\mathsf{CK}}(\boldsymbol{x}, \boldsymbol{y}) = \xi_{q,0} + \sum_{\ell=1}^{q} \sum_{S \in \mathcal{E}_\ell} \frac{r(S)\xi_{q,\ell}}{d} \cdot Y_S(\boldsymbol{x}) Y_S(\boldsymbol{y}) \,, \tag{4}$$

where $r(S) = q + 1 - \gamma(S)$ and $\xi_{q,\ell} \geq 0$ is defined in Eq. (2)*.*

Notice that Y_S with $\gamma(S) > q$ (monomials with support not contained in a segment of size q) are in the null space of H^{CK}. Hence (as long as $\xi_{q,\ell} > 0$ for all $0 \leq \ell \leq q$), the RKHS associated to H^{CK} exactly contains all the functions in the q-local function class $L^2(\mathscr{Q}^d, \text{Loc}_q)$ (c.f. Eq. (LOC)). In words, $L^2(\mathscr{Q}^d, \text{Loc}_q)$ consists of functions that are localized on patches, with no long-range interactions between different parts of the image. An example of local function with $q = 3$ is given by $f(\boldsymbol{x}) = x_1x_2x_3 + x_4x_6 + x_5$.

On the other hand, the RKHS associated to the fully-connected kernel H^{FC} (FC) typically contains all the functions in $L^2(\mathscr{Q}^d)$ (under genericity assumptions on h). The RKHS with convolution $\dim(L^2(\mathscr{Q}^d, \text{Loc}_q)) = d2^{q-1} + 1$ is significantly smaller than $\dim(L^2(\mathscr{Q}^d)) = 2^d$, which prompts the following question: *what is the statistical advantage of using H^{CK} over H^{FC} when learning functions in $L^2(\mathscr{Q}^d, \text{Loc}_q)$?*

We first consider the classical approach to bounding the test error of [3, 12, 47] which relies on the following two standard assumptions:

(A1) *Capacity condition:* we assume $\mathcal{N}(h, \lambda) := \text{Tr}[h/(h + \lambda \mathbf{I})^{-1}] \leq C_h \lambda^{-1/\alpha}$ with[2] $\alpha > 1$.

(A2) *Source condition:* $\|h^{-\beta/2} g\|_{L^2} \leq B$ with[3] $\beta > \frac{\alpha-1}{\alpha}$ and $B \geq 0$.

The capacity condition (A1) characterizes the size of the RKHS: for increasing α, the RKHS contains less and less functions. The source condition (A2) characterizes the regularity of the target function (the 'source') with respect to the kernel: increasing β corresponds to smoother and smoother functions. See Appendix B.2 for more discussions.

Based on these two assumptions, we can apply standard bounds on the KRR test error and obtain:

Theorem 1 (Generalization error of KRR with H^{CK})**.** *Let $h : \mathbb{R} \to \mathbb{R}$ be an inner-product kernel satisfying (A1). Let $f_\star \in L^2(\mathscr{Q}^d, \text{Loc}_q)$ with $f_\star(\boldsymbol{x}) = \sum_{k \in [d]} g_k(\boldsymbol{x}_{(k)})$ satisfying source condition $\sum_{k \in [d]} \|h^{-\beta/2} g_k\|_{L^2}^2 \leq B^2$. Then there exists $C_1, C_2, C_3 > 0$ constants that only depend on (A1) and B (and independent of d), such that for $n \geq C_1 \max(\|f_\star\|_{L^\infty}^2, d)$ and $\lambda_* = \frac{C_2}{d}(d/n)^{\frac{\alpha}{\alpha\beta+1}}$,*

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\big\{R(f_\star, \hat{f}_{\lambda_\star})\big\} \leq C_3 \left(\frac{d}{n}\right)^{\frac{\alpha\beta}{\alpha\beta+1}} \,. \tag{5}$$

Note that the exponent $\frac{\beta\alpha}{\beta\alpha+1}$ only depends on the q-dimensional kernel h. Hence, the generalization bound with respect to (n/d) is independent of the dimension d of the image. Let's compare to KRR with inner-product kernel H^{FC} (FC): from [12], we have the minmax rate $\mathbb{E}_{\boldsymbol{\varepsilon}}\{R(f_\star, \hat{f}_\lambda)\} \asymp n^{-\frac{\tilde{\alpha}\tilde{\beta}}{\tilde{\alpha}\tilde{\beta}+1}}$ where h is now defined in d dimension and verifies (A1) and (A2) with constants $\tilde{\alpha}, \tilde{\beta}$. Typically, if $f_\star$ is only assumed Lipschitz, then $\tilde{\beta}\tilde{\alpha} = O(1/d)$, which leads to a minmax rate $n^{-O(1/d)}$ for H^{FC}, while for H^{CK}, $\beta\alpha = O(1/q)$, which leads to a minmax rate $n^{-O(1/q)}$. Hence, for $q \ll d$,

[2] Here, h is the integral operator and $\text{Tr}[h/(h + \lambda \mathbf{I})^{-1}] = \sum_{j \geq 1} \frac{\lambda_j}{\lambda_j + \lambda}$ with $\{\lambda_j\}_{j \geq 1}$ eigenvalues of h.

[3] Again, h is the operator with $h^{-\kappa} g = \sum_{j \geq 1} \lambda_j^{-\kappa} \langle f, \psi_j \rangle \psi_j$, where $\{\psi_j\}_{j \geq 1}$ are the eigenvectors of h.

H^{CK} breaks the curse of dimensionality by restricting the RKHS to 'local' functions. Similarly, [22] derived a decay rates in n that do not depend on d for a one-layer convolutional kernel. The key difference between Theorem 1 and [22] is that we obtain a non-asymptotic bound that is minmax optimal up to a constant multiplicative factor in both d and n (this can be showed for example by adapting the proof in Appendix B.6 in [7]) using a rigorous framework of source and capacity condition.

Theorem 1 and results of this type suffers from several limitations: 1) they are tight only in a minmax sense; 2) they do not provide comparisons for specific subclasses of functions; 3) in order to obtain the minmax rate, the regularization parameter λ has to be carefully tuned to balance the bias and variance terms, which is in contrast to modern practice where often the model is trained until interpolation. This led several groups to consider instead the test error of KRR in a high-dimensional limit [11, 24, 38] and derive exact asymptotic predictions correct up to an additive vanishing constant for any $f_\star \in L^2$ (see Appendix C for more details).

Using the general framework in [38], we get the following result for q, d large:

Theorem 2 (Generalization error of KRR with H^{CK} in high-dimension (informal))**.** *Let $f_\star \in L^2(\mathscr{Q}^d, \mathrm{Loc}_q)$ and $h : \mathbb{R} \to \mathbb{R}$ verifying some 'genericity condition'. Then for $n = dq^{\mathsf{s}-1+\nu}$ with $0 < \nu < 1$, and $\lambda = O(1)$ (in particular $\lambda = 0$ works), we have*

$$\hat{f}_\lambda = \mathsf{P}_{\mathcal{E}_{\leq \mathsf{s},\nu}} f_\star + o_q(1)\,, \tag{6}$$

where $\mathsf{P}_{\mathcal{E}_{\leq \mathsf{s},\nu}}$ is the projection on the span of Y_S with either $|S| < \mathsf{s}$ and $S \in \mathcal{E}_{|S|}$ or $|S| = \mathsf{s}$ and $\gamma(S) \leq q(1 - q^{-\nu})$.

See Appendix C.1 for a rigorous statement. In words, when $dq^{\mathsf{s}-1} \ll n \ll dq^{\mathsf{s}}$, KRR with H^{CK} only learns a degree-s polynomial approximation to $f_\star$.

On the other hand, when considering the standard inner-product kernel H^{FC} (FC) we get:

Theorem 3 (Generalization error of KRR with H^{FC} in high-dimension (informal))**.** *Let $f_\star \in L^2(\mathscr{Q}^d)$ and $\tilde{h} : \mathbb{R} \to \mathbb{R}$ with some 'genericity condition'. Then for $d^{\mathsf{s}} \ll n \ll d^{\mathsf{s}+1}$ and $\lambda = O(1)$,*

$$\hat{f}_\lambda = \mathsf{P}_{\leq \mathsf{s}} f_\star + o_d(1)\,, \tag{7}$$

where $\mathsf{P}_{\leq \mathsf{s}}$ is the projection on the subspace of degree-s polynomials.

This theorem was proved in [24, 38]. Notice that Eq. (7) does not depend on the structure of $f_\star$. Hence, when $f_\star \in L^2(\mathscr{Q}^d, \mathrm{Loc}_q)$, Theorems 2 and 3 shows a clear statistical advantage of H^{CK} over H^{FC} when $q \ll d$ (and therefore of one-layer CNNs over fully-connected neural networks in the kernel regime).

2.3 Local average pooling and downsampling

In many applications such as object recognition, we expect the target function to depend mildly on the absolute spatial position of an object and to be stable under small shifts of the input. To take this local invariance into account, convolution layers are often followed by a pooling operation. Here we consider local average pooling on a segment of length ω and obtain the kernel

$$H^{\mathsf{CK}}_\omega(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{d\omega} \sum_{k \in [d]} \sum_{s, s' \in [\omega]} h\left(\langle \boldsymbol{x}_{(k+s)}, \boldsymbol{y}_{(k+s')} \rangle / q \right)\,. \tag{8}$$

Define $\mathcal{S}_\ell = \{S \subseteq [q] : |S| = \ell\}$ as the collection of sets of size ℓ. We further define an equivalence relation $\sim$ on $\mathcal{S}_\ell$: $S \sim S'$ if S' is a translated subset of S in $[q]$ (without cyclic convention). We denote $\mathcal{C}_\ell$ the quotient set of $\mathcal{A}_\ell$ under the equivalence relation $\sim$.

Proposition 2 (Eigendecomposition of H^{CK}_ω)**.** *Let H^{CK}_ω be a convolutional kernel with local average pooling as defined in Eq.* (8)*. Then H^{CK}_ω admits the following eigendecomposition:*

$$H^{\mathsf{CK}}_\omega(\boldsymbol{x}, \boldsymbol{y}) = \omega \xi_{q,0} + \sum_{\ell=1}^{q} \sum_{S \in \mathcal{C}_\ell} \sum_{j \in [d]} \frac{\kappa_j r(S) \xi_{q,\ell}}{d} \cdot \psi_{j,S}(\boldsymbol{x}) \psi_{j,S}(\boldsymbol{y})\,, \tag{9}$$

where (denoting $k + S$ the translated set S by k positions with cyclic convention in $[d]$)

$$\kappa_j = 1 + 2 \sum_{k=1}^{\omega-1} (1 - k/\omega) \cos\left(\frac{2\pi j k}{d} \right)\,, \qquad \psi_{j,S}(\boldsymbol{x}) = \frac{1}{\sqrt{d}} \sum_{k=1}^{d} e^{\frac{2i\pi j k}{d}} Y_{k+S}(\boldsymbol{x})\,. \tag{10}$$

First notice that, as long as $\mathsf{gcd}(\omega, d) = 1$, the RKHS associated to H^{CK} contains the same set of functions as the RKHS of H^{CK}, i.e., all local functions $L^2(\mathscr{Q}^d, \mathrm{Loc}_q)$. (There are $\mathsf{gcd}(\omega, d) - 1$ number of zero weights: $\kappa_j = 0$ for all $j \in [d-1]$ such that d is a divisor of $j\omega$. See Appendix A.3 for details.) However H^{CK} will penalize different frequency components of the functions differently. Denote $f_j(\boldsymbol{x})$ the j-th component of the discrete Fourier transform of the function, i.e., $f_j(\boldsymbol{x}) = \frac{1}{\sqrt{d}} \sum_{k\in[d]} \rho_j^k f(t_k \cdot \boldsymbol{x})$ where $\rho_j = e^{2i\pi j/d}$ and $t_k \cdot \boldsymbol{x} = (x_{k+1}, \dots, x_d, x_1, \dots, x_k)$ is the cyclic shift by k pixels. Then H^{CK} reweights the eigenspaces associated with $f_j(\boldsymbol{x})$ by a factor κ_j, promoting low-frequency components ($\kappa_j > 1$) and penalizing the high-frequencies ($\kappa_j < 1$). In words, *pooling biases the learning towards low-frequency functions*, which are stable by small shifts.

Let us focus on two special choices here: the pooling parameter $\omega = 1$ and $\omega = d$. When $\omega = 1$, H^{CK}_ω reduces to H^{CK} ($\kappa_j = 1$ for all $j \in [d]$) which does not bias towards either low or high frequency components. When $\omega = d$, we denote such kernel $H^{\mathsf{CK}}_{\omega=d}$ by $H^{\mathsf{CK}}_{\mathsf{GP}}$ which corresponds to global average pooling. In this case, we have $\kappa_d = d$ and $\kappa_j = 0$ for $j < d$ which enforces exact invariance under the group of cyclic translations. More precisely, $H^{\mathsf{CK}}_{\mathsf{GP}}$ has RKHS that contains all cyclic q-local functions $f(\boldsymbol{x}) = \sum_{k\in[d]} g(\boldsymbol{x}_{(k)}) \in L^2(\mathscr{Q}^d, \mathrm{CycLoc}_q)$ (c.f. Eq. (CYC-LOC)).

We obtain a bound on the test error of KRR with H^{CK}_ω similar to Theorem 1, but with d replaced by an effective dimension d^{eff}.

Theorem 4 (Generalization of KRR with average pooling (fixed d, q))**.** *Assume that $h : \mathbb{R} \to \mathbb{R}$ has $\xi_{q,0} = 0$ and satisfies (A1). Further assume $(A2')$ that $\|(H^{\mathsf{CK}}_\omega/\omega)^{-\beta/2} f_\star\|_{L^2} \le B$. Define $d^{\text{eff}} = \sum_{j\in[d]:\kappa_j>0} (\kappa_j/\omega)^{1/\alpha}$. Then there exists $C_1, C_2, C_3 > 0$ constants independent of d, such that for $n \ge C_1 \max(\|f_\star\|^2_{L^\infty}, d_{\text{eff}})$ and setting $\lambda_* = C_2 (d_{\text{eff}}/n)^{\frac{\alpha}{\alpha\beta+1}}$, we get*

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\big\{R(f_\star, \hat{f}_{\lambda_\star})\big\} \le C_3 \left(\frac{d_{\text{eff}}}{n}\right)^{\frac{\alpha\beta}{\alpha\beta+1}} . \tag{11}$$

By Jensen's inequality, we have $d^{\text{eff}} \le d/\omega^{1/\alpha}$. In particular, for global pooling, $d^{\text{eff}} = 1$ and the bound (11) does not depend on d at all. Adding average pooling improve by a factor $\omega^{1/\alpha}$ the upper bound on the sample complexity for fitting low-frequency functions. Can we confirm this statistical advantage using the predictions for KRR in high dimension? Consider first the case of global pooling:

Theorem 5 (Generalization of KRR with $H^{\mathsf{CK}}_{\mathsf{GP}}$ in high-dimension (informal))**.** *Let $f_\star \in L^2(\mathscr{Q}^d, \mathrm{CycLoc}_q)$ and $h : \mathbb{R} \to \mathbb{R}$ verifying some 'genericity condition'. Then for $n = q^{\mathsf{s}-1+\nu}$ with $0 < \nu < 1$, and $\lambda = O(1)$, we have ($\mathsf{P}_{\mathcal{E}_{\le \mathsf{s},\nu}}$ is defined as in Theorem 2)*

$$\hat{f}_\lambda = \mathsf{P}_{\mathcal{E}_{\le \mathsf{s},\nu}} f_\star + o_q(1) . \tag{12}$$

Hence, global average pooling results in an improvement by a factor d in statistical efficiency when fitting cyclic local functions, compared to H^{CK}. This improvement was already noticed in [7, 39] but in the case of $q = d$ (fully connected neural networks).

For $\omega < d$, a direct application of the theorems in [38] is more challenging because of the mixing of eigenvalues. In this case, a modification of [38], where eigenvalues are not necessary ordered anymore would apply. However, for simplicity, we present in Appendix C.1 a simplified kernel with non-overlapping local pooling which we believe captures the statistical behavior of local pooling. In this case, we show that Theorem 5 holds with $n = (d/\omega) \cdot q^{\mathsf{s}-1+\nu}$, which interpolates between Theorem 2 ($\omega = 1$) and Theorem 5 ($\omega = d$).

Downsampling: Often pooling is associated with a downsampling operation, which subsample one every Δ output coordinates. In Appendix A.4, we characterize the eigendecomposition of $H^{\mathsf{CK}}_{\omega,\Delta}$ (Proposition 4) and prove for the popular choice $\omega = \Delta$, that downsampling does not modify the cyclic invariant subspace $j = d$ (Proposition 5). More generally, we conjecture and check numerically that downsampling with $\Delta \le \omega$ leaves the low-frequency eigenspaces approximately unchanged. In particular, the statistical complexity of learning low-frequency functions is not modified by downsampling operation in the one-layer case (while downsampling is potentially beneficial in further layers).

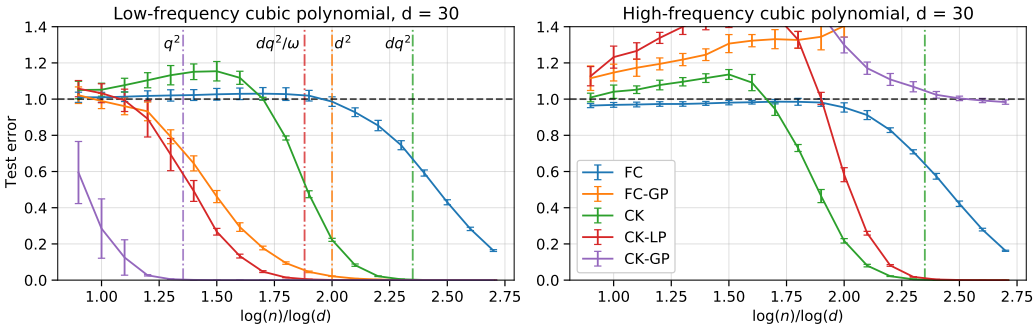

Figure 1: Learning low-frequency (left) and high-frequency (right) cubic polynomials over the hypercube $d = 30$, using KRR with H^{FC} (FC), $H^{\mathsf{FC}}_{\mathsf{GP}}$ (FC-GP), H^{CK} (CK), H^{CK}_{ω} (CK-LP) and $H^{\mathsf{CK}}_{\mathsf{GP}}$ (CK-GP), and regularization parameter $\lambda = 0^+$. We report the average and the standard deviation of the test error over 5 realizations, against the sample size n.

3 Numerical simulations

In order to check our theoretical predictions, we perform a simple numerical experiment on simulated data. We take $\boldsymbol{x} \sim \mathrm{Unif}(\mathscr{Q}^d)$ with $d = 30$, and consider two target functions:

$$f_{\mathsf{LF},3}(\boldsymbol{x}) = \frac{1}{\sqrt{d}} \sum_{i\in[d]} x_i x_{i+1} x_{i+2}\,, \qquad f_{\mathsf{HF},3}(\boldsymbol{x}) = \frac{1}{\sqrt{d}} \sum_{i\in[d]} (-1)^i \cdot x_i x_{i+1} x_{i+2}\,. \tag{13}$$

Here $f_{\mathsf{LF},3}$ is a cyclic-invariant local polynomial ($f_{\mathsf{LF},3}$ is 'low-frequency'). The function $f_{\mathsf{HF},3}$ is a high-frequency local polynomial, and is orthogonal to the space of cyclic invariant functions. On these target functions, we compare the test error of kernel ridge regression with 5 different kernels: a standard inner-product kernel $H^{\mathsf{FC}}(\boldsymbol{x},\boldsymbol{y}) = h(\langle \boldsymbol{x},\boldsymbol{y}\rangle/d)$; a cyclic invariant kernel $H^{\mathsf{FC}}_{\mathsf{GP}}(\boldsymbol{x},\boldsymbol{y})$ (convolutional kernel with global pooling and full-size patches $q = d$); a convolutional kernel H^{CK} with patch size $q = 10$; a convolutional kernel with local pooling H^{CK}_{ω} with $q = 10$ and $\omega = 5$; and a convolutional kernel with global pooling $H^{\mathsf{CK}}_{\mathsf{GP}}$ with $q = 10$. In all these kernels, we choose a common $h(t) = \sum_{i\in[5]} 0.2 * t^i$ which is a degree 5-polynomial.

In Figure 1, we report the test errors of fitting $f_{\mathsf{LF},3}$ (left) and $f_{\mathsf{HF},3}$ (right) using kernel ridge regression with these 5 kernels. We choose a small regularization parameter $\lambda = 10^{-6}$, and the noise level $\sigma_\varepsilon = 0$. The curves are averaged over 5 independent instances and the error bar stands for the standard deviation of these instances. The results match well our theoretical predictions. For the function $f_{\mathsf{LF},3}$, the sample sizes required to achieve vanishing test errors are ordered as $H^{\mathsf{CK}}_{\mathsf{GP}} < H^{\mathsf{CK}}_{\omega} < H^{\mathsf{CK}} < H^{\mathsf{FC}}_{\mathsf{GP}} < H^{\mathsf{FC}}$ and are around the predicted thresholds $q^2 < dq^2/\omega < d^2 < dq^2 < d^3$ respectively. Next we look at the test error of fitting the high frequency local function $f_{\mathsf{HF},3}$. The test errors of H^{CK} and H^{FC} are the same for $f_{\mathsf{HF},3}$ and $f_{\mathsf{LF},3}$: this is because these kernels do not have bias towards either high-frequency or low-frequency functions. The kernel H^{CK}_{ω} perform worse on $f_{\mathsf{HF},3}$ than on $f_{\mathsf{LF},3}$: this is because the eigenspaces of H^{CK}_{ω} are biased towards low-frequency polynomials. The kernels $H^{\mathsf{CK}}_{\mathsf{GP}}$ and $H^{\mathsf{FC}}_{\mathsf{GP}}$ do not fit $f_{\mathsf{HF},3}$ at all (test error greater than or equal to 1): this is because the RKHS of these two kernels only contain cyclic polynomials, but $f_{\mathsf{HF},3}$ is orthogonal to the space of cyclic polynomials.

4 Discussion and Future Work

In this paper, we characterized in a stylized setting how convolution, average pooling and downsampling operations modify the RKHS, by restricting it to q-local functions and then biasing the RKHS towards low-frequency components. We quantified precisely the gain in statistical efficiency of KRR using these operations. Beyond illustrating the 'RKHS engineering' of image-like function classes, these results can further provide intuition and a rigorous foundation for convolution and pooling operations in kernels and CNNs. A natural extension would be to study the multilayer convolutional kernels in details and consider other pooling operations such as max-pooling. Another important question is how anisotropy of the data impacts the results of this paper: in particular, it was shown that pre-processing (whitening of the patches) greatly improves the performance of convolutional kernels [6, 46]. A more challenging question is to study how training and feature learning can further improve the performance of CNNs outside the kernel regime.

Acknowledgement

S.M. is supported by NSF grant DMS-2210827. T.M. is supported by NSF through award DMS-2031883 and the Simons Foundation through Award 814639 for the Collaboration on the Theoretical Foundations of Deep Learning. T.M. also acknowledge the NSF grant CCF-2006489 and the ONR grant N00014-18-1-2729.

References

[1] Z. Allen-Zhu, Y. Li, and Z. Song. On the convergence rate of training recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 6676–6688. 2019.

[2] S. Arora, S. S. Du, W. Hu, Z. Li, R. Salakhutdinov, and R. Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, pages 8139–8148, 2019.

[3] F. Bach. *Learning Theory from First Principles*. 2021.

[4] W. Beckner. Inequalities in Fourier analysis. *Annals of Mathematics*, pages 159–182, 1975.

[5] W. Beckner. Sobolev inequalities, the Poisson semigroup, and analysis on the sphere S^n. *Proceedings of the National Academy of Sciences*, 89(11):4816–4819, 1992.

[6] A. Bietti. Approximation and learning with deep convolutional models: a kernel perspective. *arXiv preprint arXiv:2102.10032*, 2021.

[7] A. Bietti, L. Venturi, and J. Bruna. On the sample complexity of learning with geometric stability. *arXiv preprint arXiv:2106.07148*, 2021.

[8] A. Bonami. Etude des coefficients de Fourier des fonctions de $L^p(G)$. In *Annales de l'institut Fourier*, volume 20, pages 335–402, 1970.

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

[10] Y.-L. Boureau, J. Ponce, and Y. LeCun. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 111–118, 2010.

[11] A. Canatar, B. Bordelon, and C. Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature communications*, 12(1):1–12, 2021.

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

[13] L. Chizat, E. Oyallon, and F. Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2933–2943, 2019.

[14] A. Coates, A. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. In G. Gordon, D. Dunson, and M. Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 215–223, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

[15] N. Cohen and A. Shashua. Convolutional rectifier networks as generalized tensor decompositions. In *International Conference on Machine Learning*, pages 955–963. PMLR, 2016.

[16] N. Cohen and A. Shashua. Inductive bias of deep convolutional networks through pooling geometry. *arXiv preprint arXiv:1605.06743*, 2016.

[17] H. Cui, B. Loureiro, F. Krzakala, and L. Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *arXiv preprint arXiv:2105.15004*, 2021.

[18] A. Daniely, R. Frostig, and Y. Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances in Neural Information Processing Systems*, pages 2253–2261, 2016.

[19] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai. Gradient descent finds global minima of deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1675–1685. PMLR, 09–15 Jun 2019.

[20] S. S. Du, X. Zhai, B. Poczos, and A. Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=S1eK3i09YQ`.

[21] N. El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, 2010.

[22] A. Favero, F. Cagnetta, and M. Wyart. Locality defeats the curse of dimensionality in convolutional teacher-student scenarios. *arXiv preprint arXiv:2106.08619*, 2021.

[23] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. When do neural networks outperform kernel methods? *Advances in Neural Information Processing Systems*, 33, 2020.

[24] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021.

[25] L. Gross. Logarithmic sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

[26] Z. Harchaoui, F. R. Bach, and E. Moulines. Testing for homogeneity with kernel fisher discriminant analysis. In *NIPS*, pages 609–616. Citeseer, 2007.

[27] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[28] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8580–8589, 2018.

[29] A. Jacot, B. Şimşek, F. Spadaro, C. Hongler, and F. Gabriel. Kernel alignment risk estimator: Risk prediction from training data. *arXiv preprint arXiv:2006.09796*, 2020.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[31] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[32] Z. Li, R. Wang, D. Yu, S. S. Du, W. Hu, R. Salakhutdinov, and S. Arora. Enhanced convolutional neural tangent kernels. *arXiv preprint arXiv:1911.00809*, 2019.

[33] Z. Li, Y. Zhang, and S. Arora. Why are convolutional nets more sample-efficient than fully-connected nets? *arXiv preprint arXiv:2010.08515*, 2020.

[34] T. Liang, A. Rakhlin, et al. Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48(3):1329–1347, 2020.

[35] J. Mairal. End-to-end kernel learning with supervised convolutional kernel networks. *arXiv preprint arXiv:1605.06265*, 2016.

[36] J. Mairal, P. Koniusz, Z. Harchaoui, and C. Schmid. Convolutional kernel networks. *arXiv preprint arXiv:1406.3332*, 2014.

[37] E. Malach and S. Shalev-Shwartz. Computational separation between convolutional and fully-connected networks. *arXiv preprint arXiv:2010.01369*, 2020.

[38] S. Mei, T. Misiakiewicz, and A. Montanari. Generalization error of random features and kernel methods: hypercontractivity and kernel matrix concentration. *arXiv preprint arXiv:2101.10588*, 2021.

[39] S. Mei, T. Misiakiewicz, and A. Montanari. Learning with invariances in random features and kernel models. *arXiv preprint arXiv:2102.13219*, 2021.

[40] H. N. Mhaskar and T. Poggio. Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications*, 14(06):829–848, 2016.

[41] R. O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

[42] M. Rudelson and R. Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

[43] M. Scetbon and Z. Harchaoui. Harmonic decompositions of convolutional networks. In *International Conference on Machine Learning*, pages 8522–8532. PMLR, 2020.

[44] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[45] V. Shankar, A. Fang, W. Guo, S. Fridovich-Keil, J. Ragan-Kelley, L. Schmidt, and B. Recht. Neural kernels without tangents. In *International Conference on Machine Learning*, pages 8614–8623. PMLR, 2020.

[46] L. Thiry, M. Arbel, E. Belilovsky, and E. Oyallon. The unreasonable effectiveness of patches in deep convolutional kernels methods. *arXiv preprint arXiv:2101.07528*, 2021.

[47] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[48] L. Xiao. Eigenspace restructuring: a principle of space and frequency in neural networks. *arXiv preprint arXiv:2112.05611*, 2021.

[49] D. Zou, Y. Cao, D. Zhou, and Q. Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv:1811.08888*, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] All mathematical statements are proved in the Appendix.
   (b) Did you describe the limitations of your work? [Yes] See Sections 1.1 and 4.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This is a theoretical work with no societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See main text and Appendix.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A] Our paper is theoretical and we only provide small numerical illustrations (which are fully mathematically explicit and only require matrix inversions).
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Contents

1 Introduction **1**

1.1 Summary of main results . . . 3

1.2 Related work . . . 4

2 Main results **5**

2.1 Functions on the hypercube and eigendecomposition of kernel operators . . . 5

2.2 One-layer convolutional kernel . . . 5

2.3 Local average pooling and downsampling . . . 7

3 Numerical simulations **9**

4 Discussion and Future Work **9**

A Details from the main text **15**

A.1 Notations . . . 15

A.2 Convolutional neural tangent kernel . . . 15

A.3 Local average pooling operation . . . 16

A.4 Downsampling operation . . . 17

A.5 Multilayer convolutional kernels . . . 21

A.6 Proofs diagonalization of convolutional kernels . . . 22

A.7 Additional numerical simulations . . . 24

B Generalization error of kernel methods in fixed dimension **26**

B.1 Bound on kernel methods using Rademacher complexities . . . 26

B.2 Generalization error of KRR in the classical regime . . . 27

C Generalization error of KRR in high dimension **29**

C.1 Generalization error of convolutional kernels in high dimension . . . 30

C.2 Checking the assumptions . . . 33

C.3 Proof of Proposition 6 . . . 34

C.4 Proof of Theorem 7 . . . 36

C.5 Proof of Theorem 8 . . . 38

C.6 Auxiliary results . . . 40

D Technical background of function spaces on the hypercube **43**

D.1 Fourier basis . . . 44

D.2 Hypercubic Gegenbauer . . . 44

D.3 Hermite polynomials . . . 45

D.4 Hypercontractivity of uniform distributions on the hypercube . . . 45

A Details from the main text

A.1 Notations

For a positive integer, we denote by $[n]$ the set $\{1, 2, \dots, n\}$. For vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, we denote $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = u_1 v_1 + \dots + u_d v_d$ their scalar product, and $\|\boldsymbol{u}\|_2 = \langle \boldsymbol{u}, \boldsymbol{u} \rangle^{1/2}$ the ℓ_2 norm. Given a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times m}$, we denote $\|\boldsymbol{A}\|_{\mathrm{op}} = \max_{\|\boldsymbol{u}\|_2 = 1} \|\boldsymbol{A}\boldsymbol{u}\|_2$ its operator norm and by $\|\boldsymbol{A}\|_F = \big(\sum_{i,j} A_{ij}^2\big)^{1/2}$ its Frobenius norm. If $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is a square matrix, the trace of $\boldsymbol{A}$ is denoted by $\mathrm{Tr}(\boldsymbol{A}) = \sum_{i \in [n]} A_{ii}$.

We use $O_d(\,\cdot\,)$ (resp. $o_d(\,\cdot\,)$) for the standard big-O (resp. little-o) relations, where the subscript d emphasizes the asymptotic variable. Furthermore, we write $f = \Omega_d(g)$ if $g(d) = O_d(f(d))$, and $f = \omega_d(g)$ if $g(d) = o_d(f(d))$. Finally, $f = \Theta_d(g)$ if we have both $f = O_d(g)$ and $f = \Omega_d(g)$.

We use $O_{d,\mathbb{P}}(\,\cdot\,)$ (resp. $o_{d,\mathbb{P}}(\,\cdot\,)$) the big-O (resp. little-o) in probability relations. Namely, for $h_1(d)$ and $h_2(d)$ two sequences of random variables, $h_1(d) = O_{d,\mathbb{P}}(h_2(d))$ if for any $\varepsilon > 0$, there exists $C_\varepsilon > 0$ and $d_\varepsilon \in \mathbb{Z}_{>0}$, such that

$$\mathbb{P}(|h_1(d)/h_2(d)| > C_\varepsilon) \le \varepsilon, \qquad \forall d \ge d_\varepsilon,$$

and respectively: $h_1(d) = o_{d,\mathbb{P}}(h_2(d))$, if $h_1(d)/h_2(d)$ converges to 0 in probability. Similarly, we will denote $h_1(d) = \Omega_{d,\mathbb{P}}(h_2(d))$ if $h_2(d) = O_{d,\mathbb{P}}(h_1(d))$, and $h_1(d) = \omega_{d,\mathbb{P}}(h_2(d))$ if $h_2(d) = o_{d,\mathbb{P}}(h_1(d))$. Finally, $h_1(d) = \Theta_{d,\mathbb{P}}(h_2(d))$ if we have both $h_1(d) = O_{d,\mathbb{P}}(h_2(d))$ and $h_1(d) = \Omega_{d,\mathbb{P}}(h_2(d))$.

A.2 Convolutional neural tangent kernel

In this section, we justify the expression of the convolutional neural tangent kernel $H^{\mathsf{CK}}_{\boldsymbol{w},\Delta}$ (CK-AP-DS), obtained as the tangent kernel of a neural network composed of a one convolution layer followed by local average pooling and downsampling (CNN-AP-DS).

Proposition 3. *Let $\sigma \in \mathcal{C}^1(\mathbb{R})$ be an activation function. Consider the following one-layer convolutional neural network with ω-local average pooling and Δ-downsampling:*

$$f_N^{\mathsf{CNN}}(\boldsymbol{x}; \boldsymbol{\Theta}) = \sum_{i \in [N]} \sum_{k \in [d/\Delta]} a_{ik} \sum_{s \in [\omega]} \sigma\big(\langle \boldsymbol{w}_i, \boldsymbol{x}_{(k\Delta+s)} \rangle\big)\,. \tag{14}$$

Let $a_{ik}^0 \sim_{\text{i.i.d.}} \mathsf{N}(0,1)$ and $\sqrt{q}\boldsymbol{w}_i^0 \sim_{\text{i.i.d.}} \mathrm{Unif}(\mathscr{Q}^q)$ independently, and $\boldsymbol{\Theta}^0 = \{(a_{ik}^0)_{i \in [N], k \in [d/\Delta]}, (\boldsymbol{w}_i^0)_{i \in [N]}\}$. Then there exists $h : [-1,1] \to \mathbb{R}$, such that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathscr{Q}^d$, we have almost surely

$$\lim_{N \to \infty} \big\langle \nabla_{\boldsymbol{\Theta}} f_N^{\mathsf{CNN}}(\boldsymbol{x}; \boldsymbol{\Theta}^0), \nabla_{\boldsymbol{\Theta}} f_N^{\mathsf{CNN}}(\boldsymbol{y}; \boldsymbol{\Theta}^0) \big\rangle / N = \sum_{k \in [d/\Delta]} \sum_{s, s' \in [\omega]} h\big(\langle \boldsymbol{x}_{(k\Delta+s)}, \boldsymbol{y}_{(k\Delta+s')} \rangle / q\big)\,. \tag{15}$$

Proof of Proposition 3. For $\boldsymbol{u}, \boldsymbol{v} \in \mathscr{Q}^q$, define

$$h^{(1)}(\langle \boldsymbol{u}, \boldsymbol{v} \rangle / q) = \mathbb{E}_{\boldsymbol{w} \sim \mathrm{Unif}(\mathscr{Q}^q)} \big[\sigma(\langle \boldsymbol{u}, \boldsymbol{w} \rangle / \sqrt{q}) \sigma(\langle \boldsymbol{v}, \boldsymbol{w} \rangle / \sqrt{q})\big]\,,$$
$$h^{(2)}(\langle \boldsymbol{u}, \boldsymbol{v} \rangle / q) = \mathbb{E}_{\boldsymbol{w} \sim \mathrm{Unif}(\mathscr{Q}^q)} \big[\sigma'(\langle \boldsymbol{u}, \boldsymbol{w} \rangle / \sqrt{q}) \sigma'(\langle \boldsymbol{v}, \boldsymbol{w} \rangle / \sqrt{q}) \langle \boldsymbol{u}, \boldsymbol{v} \rangle\big] / q\,.$$

The functions $h^{(1)}, h^{(2)}$ are well defined (the RHS only depend on the inner product $\langle \boldsymbol{u}, \boldsymbol{v} \rangle$) and can be extended to functions $h^{(1)}, h^{(2)} : [-1,1] \to \mathbb{R}$.

Computing the derivative of the convolutional neural network with respect to $\boldsymbol{a} = (a_{ik}^0)_{i \in [N], k \in [d/\Delta]}$, we have

$$\begin{aligned} &\frac{1}{N} \big\langle \nabla_{\boldsymbol{a}} f_N^{\mathsf{CNN}}(\boldsymbol{x}; \boldsymbol{\Theta}^0), \nabla_{\boldsymbol{a}} f_N^{\mathsf{CNN}}(\boldsymbol{y}; \boldsymbol{\Theta}^0) \big\rangle \\ &= \sum_{k \in [d/\Delta]} \sum_{s, s' \in [\omega]} \frac{1}{N} \sum_{i \in [N]} \sigma\big(\langle \boldsymbol{w}_i^0, \boldsymbol{x}_{(k\Delta+s)} \rangle\big) \sigma\big(\langle \boldsymbol{w}_i^0, \boldsymbol{x}_{(k\Delta+s')} \rangle\big)\,. \end{aligned}$$

Hence by law of large number, we have almost surely

$$\lim_{N \to \infty} \frac{1}{N} \big\langle \nabla_{\boldsymbol{a}} f_N^{\mathsf{CNN}}(\boldsymbol{x}; \boldsymbol{\Theta}^0), \nabla_{\boldsymbol{a}} f_N^{\mathsf{CNN}}(\boldsymbol{y}; \boldsymbol{\Theta}^0) \big\rangle = \sum_{k \in [d/\Delta]} \sum_{s, s' \in [\omega]} h^{(1)}\big(\langle \boldsymbol{x}_{(k\Delta+s)}, \boldsymbol{y}_{(k\Delta+s')} \rangle / q\big)\,.$$

Similarly, computing the derivative with respect to $\sqrt{q}\boldsymbol{W} = (\sqrt{q}\boldsymbol{w}_i^0)_{i\in[N]}$ gives

$$\frac{1}{N}\langle \nabla_{\boldsymbol{W}} f_N^{\mathsf{CNN}}(\boldsymbol{x};\boldsymbol{\Theta}^0), \nabla_{\boldsymbol{W}} f_N^{\mathsf{CNN}}(\boldsymbol{y};\boldsymbol{\Theta}^0)\rangle$$
$$= \sum_{k,k'\in[d/\Delta]} \sum_{s,s'\in[\omega]} \frac{1}{N} \sum_{i\in[N]} a_{ik}a_{ik'}\sigma'\big(\langle \boldsymbol{w}_i^0, \boldsymbol{x}_{(k\Delta+s)}\rangle\big)\sigma'\big(\langle \boldsymbol{w}_i^0, \boldsymbol{x}_{(k'\Delta+s')}\rangle\big)\frac{\langle \boldsymbol{x}_{(k\Delta+s)}, \boldsymbol{x}_{(k'\Delta+s')}\rangle}{q}\,.$$

By law of large number, using that a_{ik} and $a_{ik'}$ are independent of mean zero and variance 1, we get almost surely

$$\lim_{N\to\infty} \frac{1}{N}\langle \nabla_{\boldsymbol{W}} f_N^{\mathsf{CNN}}(\boldsymbol{x};\boldsymbol{\Theta}^0), \nabla_{\boldsymbol{W}} f_N^{\mathsf{CNN}}(\boldsymbol{y};\boldsymbol{\Theta}^0)\rangle = \sum_{k\in[d/\Delta]} \sum_{s,s'\in[\omega]} h^{(2)}\big(\langle \boldsymbol{x}_{(k\Delta+s)}, \boldsymbol{y}_{(k\Delta+s')}\rangle/q\big)\,.$$

Taking $h = h^{(1)} + h^{(2)}$ concludes the proof. □

A.3 Local average pooling operation

Consider a function $f \in L^2(\mathscr{Q}^d)$: we can decompose it as

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{d}} \sum_{j\in[d]} f_j(\boldsymbol{x})\,, \tag{16}$$

$$f_j(\boldsymbol{x}) = \frac{1}{\sqrt{d}} \sum_{k\in[d]} \rho_j^k f(t_k\cdot \boldsymbol{x})\,, \tag{17}$$

where $\rho_j = e^{\frac{2i\pi j}{d}}$ and $t_k \cdot \boldsymbol{x} = (x_{k+1}, \dots, x_d, x_1, \dots, x_k)$ is the cyclic shift of $\boldsymbol{x}$ by k pixels. We can think about $f_j(\boldsymbol{x})$ as the j-th component of the discrete Fourier transform of the function $f(\boldsymbol{x})$ seen as a d-dimensional vector $\{f(t_k\cdot\boldsymbol{x})\}_{k\in[d]}$ for any $\boldsymbol{x}\in\mathscr{Q}^d$.

Notice furthermore that if f is a local function, i.e., f can be decomposed as a sum of functions on patches $f(\boldsymbol{x}) = \sum_{k\in[d]} g_k(\boldsymbol{x}_{(k)})$, then we can write

$$f_j(\boldsymbol{x}) = \frac{1}{\sqrt{d}} \sum_{k\in[d]} \rho_j^k f(t_k\cdot\boldsymbol{x}) = \frac{1}{\sqrt{d}} \sum_{k,u\in[d]} \rho_j^k g_u(\boldsymbol{x}_{(u+k)}) = \frac{1}{\sqrt{d}} \sum_{k\in[d]} \rho_j^k \tilde{g}_j(\boldsymbol{x}_{(k)})\,,$$

where we denoted ($\boldsymbol{v}\in\mathscr{Q}^q$)

$$\tilde{g}_j(\boldsymbol{v}) = \sum_{u\in[d]} \rho_j^{-u} g_u(\boldsymbol{v})\,.$$

In particular, decomposing $\tilde{g}_j$ in the Fourier basis, we get (denoting $c_S = \langle \tilde{g}_j, Y_S\rangle_{L^2}$),

$$f_j(\boldsymbol{x}) = \frac{1}{\sqrt{d}} \sum_{k\in[d]} \rho_j^k \tilde{g}_j(\boldsymbol{x}_{(k)}) = \sum_{S\subseteq[q]} c_S \cdot \frac{1}{\sqrt{d}} \sum_{k\in[d]} \rho_j^k Y_{k+S}(\boldsymbol{x})\,,$$

which shows that the j-th frequency component f_j is in the span of $\{Y_{j,S}\}_{S\subseteq[q]}$. In particular, applying average pooling operation in the kernel will reweight this eigenspace by a factor κ_j.

Let us further comment on the values of κ_j. First, we have

$$\kappa_j = \sum_{k=-\omega}^{\omega} (1-k/\omega)\rho_j^k.$$

In particular, the maximal eigenvalue is attained at $j = d$ with $\kappa_d = \omega$, which corresponds to the subspace of cyclic invariant functions. Furthermore, $\kappa_j = 0$ if and only if d is a divisor of $j\omega$ for $j \le d-1$, i.e., j is a multilple of $\mathsf{gcd}(\omega, d)$. There are $\mathsf{gcd}(\omega, d) - 1$ such zero eigenvalues.

In convolutional kernels, a weighted average is often preferred to local average pooling [6, 35, 36]: in that case we consider $\tau : \mathbb{R}\to\mathbb{R}$ and obtain the kernel

$$H_\tau^{\mathsf{CK}}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{d} \sum_{k,s,s'\in[d]} \tau(d(s))\tau(d(s'))h\big(\langle \boldsymbol{x}_{(k+s)}, \boldsymbol{y}_{(k+s')}\rangle/q\big)\,,$$

where $d(s) = \min(s, d-s)$ (the distance between $k+s$ and k on $[d]$ with cyclic convention). Note that H_τ^{CK} has the same eigendecomposition as H_ω^{CK} but with different weights κ_j.

A popular choice for τ is the Gaussian filter $\tau(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$. In Figure 2, we compare the eigenvalues κ_j for local average pooling and Gaussian filter with different value of ω and σ^2. Note that the eigenvalue decay controls how much high-frequencies are penalized: faster decay induces heavier penalty on the high-frequency components.

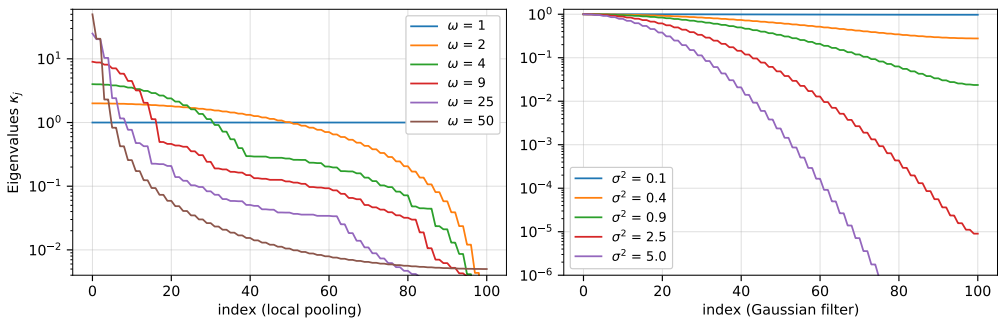

Figure 2: Decay of the weights κ_j for different lenfths ω for local average pooling (on the left) and bandwidths σ^2 for pooling with Gaussian filter (on the right), for $d = 101$.

A.4 Downsampling operation

As mentioned in the main text, a downsampling operation is often added after pooling. The kernel is given by

$$H_{\omega,\Delta}^{\mathsf{CK}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\Delta}{d\omega} \sum_{k \in [d/\Delta]} \sum_{s,s' \in [\omega]} h\left(\langle \boldsymbol{x}_{(k\Delta+s)}, \boldsymbol{y}_{(k\Delta+s')} \rangle / q \right) . \tag{18}$$

Let us introduce the family $\{\boldsymbol{M}^r\}_{r \in [q]}$ of block-circulant matrices defined by

$$M_{ij}^r = \frac{\Delta}{\omega(q+1-r)} \left| \left\{ (k,s,s',t) \in \mathcal{I}_{\omega,\Delta,r} : k\Delta + s + t \equiv i[d], k\Delta + s' + t \equiv j[d] \right\} \right| , \tag{19}$$

where we introduced the set of indices

$$\mathcal{I}_{\omega,\Delta,r} = \left\{ (k,s,s',t) : k \in [d/\Delta], s, s' \in [\omega], 0 \le t \le q-r \right\} . \tag{20}$$

We can now state the eigendecomposition of $H_{\omega,\Delta}^{\mathsf{CK}}$ in terms of the eigenvalues and eigenvectors of the matrices $\{\boldsymbol{M}^r\}_{r \in [q]}$.

Proposition 4 (Eigendecomposition of $H_{\omega,\Delta}^{\mathsf{CK}}$)**.** *Let $H_{\omega,\Delta}^{\mathsf{CK}}$ be a convolutional kernel with local average pooling and downsampling, as defined in Eq.* (18)*. Then $H_{\omega,\Delta}^{\mathsf{CK}}$ admits the following eigendecomposition:*

$$H_\omega^{\mathsf{CK}}(\boldsymbol{x}, \boldsymbol{y}) = \omega \xi_{q,0} + \sum_{\ell=1}^{q} \sum_{S \in \mathcal{C}_\ell} \sum_{j \in [d]} \frac{\xi_{q,\ell} r(S) \kappa_j^S}{d} \cdot \psi_{j,S}(\boldsymbol{x}) \psi_{j,S}(\boldsymbol{y}) , \tag{21}$$

where $\psi_{j,S}^\Delta(\boldsymbol{x}) = \sum_{k=1}^d v_{j,k}^S Y_{k+S}(\boldsymbol{x})$ with $\{\kappa_j^S, \boldsymbol{v}_j^S\}_{j \in [d]}$ eigenvalues and eigenvectors of $\boldsymbol{M}^{\gamma(S)}$.

Let us make a few comments on these matrices $\boldsymbol{M}^{\gamma(S)}$. First because they only depend on S through the diameter $\gamma(S)$, the eigenvalues and eigenvectors $\{\kappa_j^S, \boldsymbol{v}_j^S\}_{j \in [d]}$ only depend on $\gamma(S)$. Second, we see that $M_{(i+\Delta)(j+\Delta)}^{\gamma(S)} = M_{ij}^{\gamma(S)}$ and $M_{ij}^{\gamma(S)} = 0$ if $d(i,j) \ge \omega$, where $d(i,j) = \min(|i-j|, d-|i-j|)$ (i.e., the distance between i and j on the torus $[d]$). In words $\boldsymbol{M}^{\gamma(S)}$ is a

symmetric block-circulant matrix with non-zero elements on a band of size $\omega - 1$ on the left and right of the diagonal, and on the upper-right and lower-left corners. Furthermore, notice that

$$\mathrm{Tr}(\boldsymbol{M}^{\gamma(S)}) = \frac{\Delta}{d\omega r(S)} \Big| \Big\{ (k,s,t) : k \in [d/\Delta], s \in [\omega], 0 \le t \le q - \gamma(S) \Big\} \Big| = 1 \,,$$

which is independent of $\omega, \Delta, \gamma(S)$ and justify the chosen normalization. In particular, this implies that (for $\xi_{q,0} = 0$)

$$\mathrm{Tr}(\mathbb{H}^{\mathsf{CK}}_{\omega,\Delta}) := \mathbb{E}_{\boldsymbol{x}}\{H^{\mathsf{CK}}_{\omega,\Delta}(\boldsymbol{x},\boldsymbol{x})\} = \sum_{\ell \in [q]} \xi_{q,\ell} \sum_{S \in \mathcal{C}_\ell} r(S) = \sum_{\ell \in [q]} \xi_{q,\ell} B(\mathscr{Q}^q; \ell) = h(1) \,, \tag{22}$$

is also independent of the parameters (q, ω, Δ).

Example 1. Take $\Delta = 3$, $\omega = 5$, $q = 11$, then

$$\boldsymbol{M}^1 = \frac{3}{50} \left(\begin{array}{ccc|ccc|ccc|c}
18 & 15 & 11 & 7 & 4 & 0 & & & & \\
15 & 19 & 15 & 11 & 8 & 4 & 0 & & & \dots \\
11 & 15 & 18 & 14 & 11 & 7 & 3 & 0 & & \\
\hline
7 & 11 & 14 & 18 & 15 & 11 & 7 & 3 & 0 & \\
4 & 8 & 11 & 15 & 19 & 15 & 11 & 8 & 4 & \dots \\
0 & 4 & 7 & 11 & 15 & 18 & 14 & 11 & 7 & \\
\hline
 & 0 & 3 & & & & & & & \\
 & & 0 & & \vdots & & & \ddots & & \dots \\
 & & & & & & & & &
\end{array} \right) ,$$

and

$$\boldsymbol{M}^4 = \frac{3}{35} \left(\begin{array}{ccc|ccc|cc}
13 & 11 & 8 & 5 & 3 & 0 & & \\
11 & 14 & 11 & 8 & 6 & 3 & 0 & \\
8 & 11 & 13 & 10 & 8 & 5 & 2 & 0 \\
\hline
5 & 8 & 10 & & & & & \\
3 & 6 & 8 & & \ddots & & & \dots \\
0 & 3 & 5 & & & & & \\
\hline
 & \vdots & & & \vdots & & & \ddots \\
 & & & & & & &
\end{array} \right) .$$

Remark 1. Symmetric block-circulant matrices can be easily diagonalized as follows. Consider $\boldsymbol{M} = \mathsf{Circulant}(\boldsymbol{B}_1, \boldsymbol{B}_2, \dots, \boldsymbol{B}_m)$ where $\boldsymbol{B}_k \in \mathbb{R}^{\Delta \times \Delta}$, $\boldsymbol{B}_1^{\mathsf{T}} = \boldsymbol{B}_1$ and $\boldsymbol{B}_{2+k} = \boldsymbol{B}_{m-k}^{\mathsf{T}}$ for $k = 0, \dots, m-2$. Denote $\rho_j = e^{2i\pi j/m}$ and $\gamma_j(\boldsymbol{v}) = [\boldsymbol{v}, \rho_j \boldsymbol{v}, \cdots, \rho_j^{m-1}\boldsymbol{v}]/\sqrt{m} \in \mathbb{R}^{m\Delta}$ for any $\boldsymbol{v} \in \mathbb{R}^\Delta$. Introduce for $j = 0, \dots, m-1$, the matrix $\boldsymbol{H}_j \in \mathbb{R}^{\Delta \times \Delta}$ given by

$$\boldsymbol{H}_j = \boldsymbol{B}_1 + \rho_j \boldsymbol{B}_2 + \ldots + \rho_j^{m-1} \boldsymbol{B}_m \,. \tag{23}$$

The matrix $\boldsymbol{H}_j$ is Hermitian and we denote $(\lambda_{j,s})_{s \in [\Delta]}$ and $(\boldsymbol{v}_{j,s})_{s \in [\Delta]}$ its eigenvalues and eigenvectors. Then the eigenvalues and eigenvectors of $\boldsymbol{M}$ are given by $\{\lambda_{j,s}\}_{j \in [m], s \in [\Delta]}$ and $\{\gamma_j(\boldsymbol{v}_{j,s})\}_{j \in [m], s \in [\Delta]}$.

In particular, if $\Delta = 1$ and $\boldsymbol{M} = \mathsf{Circulant}(b_1, b_2, \dots, b_m)$ is a circulant matrix, then the eigenvalues are simply given by

$$\lambda_j = b_1 + \rho_j b_2 + \ldots + \rho_j^{m-1} b_m \,,$$

and eigenvectors $\boldsymbol{v}_j = [1, \rho_j, \cdots, \rho_j^{m-1}]/\sqrt{m}$.

Here we will focus on the impact of downsampling for single-layer convolutional kernels. We expect the downsampling operation to have a much more important role for the next layers: for example, increasing the scale of interactions or reducing the dimensionality of the pixel space.

We will argue below that adding a downsampling operation after local pooling leaves the low-frequency components approximately unchanged (while potentially modifying the high-frequency eigenspaces). We consider $\Delta \le \omega$: for $\Delta > \omega$, some basis functions Y_S with $S \in \mathcal{E}_\ell$ are in the null space of $H^{\mathsf{CK}}_{\omega,\Delta}$, which impact all frequencies.

To emphasize the dependency on ω, Δ, denote $\boldsymbol{M}^r_{\omega,\Delta}$ the matrix (19). We will study the change in the matrix $\boldsymbol{M}^r_{\omega,1}$ when adding downsampling Δ, and consider

$$\boldsymbol{M}^r_{\omega,\Delta} = \boldsymbol{M}^r_{\omega,1} + \boldsymbol{A}^r_{\omega,\Delta}\,, \tag{24}$$

where we denote $\boldsymbol{A}^r_{\omega,\Delta} = \boldsymbol{M}^r_{\omega,\Delta} - \boldsymbol{M}^r_{\omega,1}$. Notice that $\boldsymbol{A}^r_{\omega,\Delta}$ is a symmetric block-circulant matrix. Therefore, from Remark 1, the eigenvectors of $\boldsymbol{A}^r_{\omega,\Delta}$ are given by $\{\gamma_j(\boldsymbol{v}_{j,s})\}_{j\in[m],s\in[\Delta]}$ where $d = m\Delta$ and $\gamma_j(\boldsymbol{v}_{j,s}) = [\boldsymbol{v}_{j,s}, \rho_{m,j}\boldsymbol{v}_{j,s}, \ldots, \rho_{m,j}^{m-1}\boldsymbol{v}_{j,s}]$ with $\rho_{m,j} = e^{\frac{2i\pi j}{m}}$ and $(\boldsymbol{v}_{j,s})_{s\in[\Delta]}$ eigenvectors of $\boldsymbol{H}_j$ (23). The eigenvectors of $\boldsymbol{M}^r_{\omega,1}$ are given by $\boldsymbol{u}_t = [1, \rho_{d,t}, \cdots, \rho_{d,t}^{d-1}]/\sqrt{d}$ with $\rho_{d,t} = e^{\frac{2i\pi t}{d}}$. Notice that

$$\begin{aligned}\langle \boldsymbol{u}_t^*, \gamma_j(\boldsymbol{v}_{j,s})\rangle &= \frac{1}{\sqrt{dm}} \sum_{k\in[m]} \sum_{u\in[\Delta]} \rho_{m,j}^{k-1} \rho_{d,t}^{-(k-1)\Delta-(t-1)} (\boldsymbol{v}_{j,s})_u \\ &= \frac{1}{\sqrt{dm}} \Big(\sum_{u\in[\Delta]} \rho_{d,t}^{-(u-1)} (\boldsymbol{v}_{j,s})_u \Big) \cdot \sum_{k\in[m]} \big(\rho_{m,j}\rho_{d,t}^{-\Delta}\big)^{k-1}\,,\end{aligned}$$

which is 0 except when $t \equiv j[m]$. Hence, we see that $\boldsymbol{A}^r_{\omega,\Delta}$ in Eq. (24) only modify the eigenspaces of $\boldsymbol{M}^r_{\omega,1}$ as follows: the eigendirections $\{\gamma_j(\boldsymbol{v}_{j,s})\}_{j\in[m],s\in[\Delta]}$ coming from $\boldsymbol{H}_j$ (23) only modify the eigenspaces of $\boldsymbol{M}^r_{\omega,1}$ spanned by $\{\boldsymbol{u}_{am+j}\}_{a=0,\ldots,\Delta-1}$.

For simplicity, we will focus on the popular choice $\Delta = \omega$. Furthermore, we will only look at the impact of the eigenvalues $\boldsymbol{H}_0$ on the eigenspace spanned by $\{\boldsymbol{u}_{am}\}_{a=0,\ldots,\Delta-1}$, which contain the cyclic invariant direction. We show below that $\boldsymbol{H}_0 = \boldsymbol{0}$ and therefore $\boldsymbol{A}^r_{\omega,\omega}$ does not modify the cyclic invariant eigenspace of $\boldsymbol{M}^r_{\omega,1}$:

Proposition 5. *Consider $d = m\omega$ and the symmetric block-circulant matrix $\boldsymbol{A}^r_{\omega,\omega} = \boldsymbol{M}^r_{\omega,\omega} - \boldsymbol{M}^r_{\omega,1}$. Denote $\boldsymbol{A}^r_{\omega,\omega} = \mathsf{Circulant}(\boldsymbol{B}_1, \boldsymbol{B}_2, \ldots, \boldsymbol{B}_m)$ and*

$$\boldsymbol{H}_0 = \boldsymbol{B}_1 + \ldots + \boldsymbol{B}_m\,.$$

We have the following properties:

(a) If $q + 1 - r \equiv 0[\omega]$, then $\boldsymbol{A}^r_{\omega,\omega} = \boldsymbol{0}$, and downsampling does not modify the matrix $\boldsymbol{M}^r_{\omega,\omega} = \boldsymbol{M}^r_{\omega,1}$.

(b) We have $\boldsymbol{H}_0 = \boldsymbol{0}$ and downsampling does not modify the cyclic invariant eigenspace $\boldsymbol{A}^r_{\omega,\omega}\boldsymbol{1} = \boldsymbol{0}$.

Proof of Proposition 5. Let us first start by proving point (a). Consider $q + 1 - r = p\omega$. Fix $i \in \{0, \ldots, \Delta - 1\}$ and $\kappa \in \{0, \ldots, \omega - 1\}$. Let us compute the entry $(i, i + \kappa)$ of the matrix $\boldsymbol{M}^r_{\omega,\omega}$: this amounts to counting the number of quadruples (k, s, s', t) with $k \in [d/\omega]$, $s, s' \in [\omega]$ and $0 \le t \le p\omega - 1$, satisfying $(k\omega + s + t, k\omega + s' + t) \equiv (i, i + \kappa)[d]$. Notice that we must have $s' = s + \kappa$ and therefore $s \in \{0, \ldots, \omega - 1 - \kappa\}$. Notice that for each interval $u\omega \le t < (u + 1)\omega$ with $u \in \{0, \ldots, p - 1\}$, there are exactly $\omega - \kappa$ ways of choosing s and then t and k to satisfy the equality. We deduce that

$$(\boldsymbol{M}^r_{\omega,\omega})_{i(i+\kappa)} = \frac{\omega}{\omega(q + 1 - r)} p(\omega - \kappa) = 1 - \frac{\kappa}{\omega} = (\boldsymbol{M}^r_{\omega,1})_{i(i+\kappa)}\,.$$

By symmetry of $\boldsymbol{M}^r_{\omega,\omega}$, this concludes the proof of point (a).

Consider now point (b). First notice, because $\boldsymbol{M}^r_{\omega,\omega}$ has zero entries for $\min(|i - j|, d - |i - j|) \ge \omega$, the only non-zero blocks are $\boldsymbol{B}_1, \boldsymbol{B}_2$ and $\boldsymbol{B}_m$. Furthermore, when computing $\boldsymbol{H}_0$, the diagonal entries only have one contribution from the diagonal elements of $\boldsymbol{B}_1$. The off-diagonal elements of $\boldsymbol{H}_0$ have two contribution: one from $\boldsymbol{B}_1$ and one from $\boldsymbol{B}_2$ (if below the diagonal) or $\boldsymbol{B}_m$ (if above the diagonal), i.e.,

$$(\boldsymbol{H}_0)_{ii} = (\boldsymbol{B}_1)_{ii} \qquad (\boldsymbol{H}_0)_{i(i+\kappa)} = (\boldsymbol{B}_1)_{i(i+\kappa)} + (\boldsymbol{B}_m)_{i(i+\kappa)}\,.$$

Let us compute first the diagonal elements: we have easily, by a similar argument as above $(\boldsymbol{M}^r_{\omega,\omega})_{ii} = 1 = (\boldsymbol{M}^r_{\omega,1})_{ii}$, and therefore $\boldsymbol{H}_0$ has zero zero diagonal entries. For off-diagonal

elements, first notice that $(\boldsymbol{M}^r_{\omega,\omega})_{i(i+\kappa-\omega)} = (\boldsymbol{M}^r_{\omega,\omega})_{i(i+\omega-\kappa)}$. Then for $q+1-r = p\omega + v$, we can consider each subsegment $u\omega \leq t < (u+1)\omega$ separately, and by a simple counting argument, get $(\boldsymbol{M}^r_{\omega,\omega})_{i(i+\omega-\kappa)} + (\boldsymbol{M}^r_{\omega,\omega})_{i(i+\kappa)} = 1 - \frac{\kappa}{\omega}$. We deduce that $(\boldsymbol{H}_0)_{i(i+\kappa)} = 0$, which by symmetry implies $\boldsymbol{H}_0 = \mathbf{0}$ and concludes the proof. □

From the above result, we conjecture that more generally, for $\Delta \leq \omega$, the low-frequency eigenspaces of H^{CK}_ω remain approximately unchanged when applying a downsampling operation. We verify this conjecture numerically in several examples. In Figure 3, we plot the eigenvalues κ_j with and without downsampling. On the left, we compare κ_j for fixed $\omega = 25$ and increasing Δ. We notice that the eigenvalues do not change much for $\Delta \leq \omega$, and for $\Delta > \omega$, some κ_j become null, as discussed above. On the right, we plot κ_j for $\Delta = 1$ (continuous line) and $\Delta = \omega$ (dashed lines) for several ω. As conjectured, the top eigenvalues (low-frequency) are left approximately unchanged. In Figure 4, we plot a heatmap of the eigenvectors ordered vertically from highest associated eigenvalue (bottom) to lowest (top) for a fixed $\omega = 25$ and increasing downsampling $\Delta \in \{1, 25, 40\}$. First indeed check that the top eigenvectors correspond to low-frequency functions and the bottom eigenvectors correspond to high-frequency functions. Second, most eigenvectors are not much modified between $\Delta = 1$ and $\Delta = \omega = 25$. For the case, $\Delta > \omega$, the top eigenvectors corresponds still low-frequency functions.

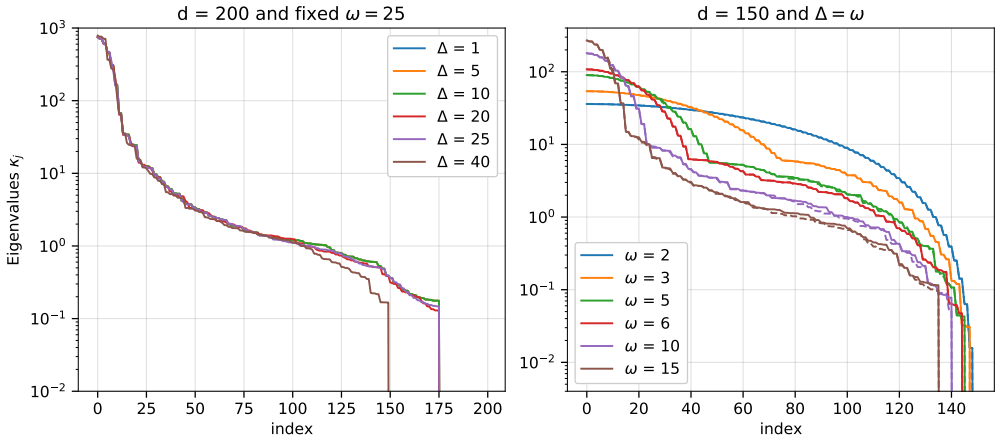

Figure 3: Impact of downsampling on the eigenvalues κ_j. On the left, we fix $\omega = 25$ ($d = 200$, $q = 30$, $r = 1$) and increase δ from 1 (no downsampling) to 40. On the right, we compare $\Delta = 1$ (continuous line) and $\Delta = \omega$ (dashed lines), with $d = 150$,$q = 20$,$r = 1$.

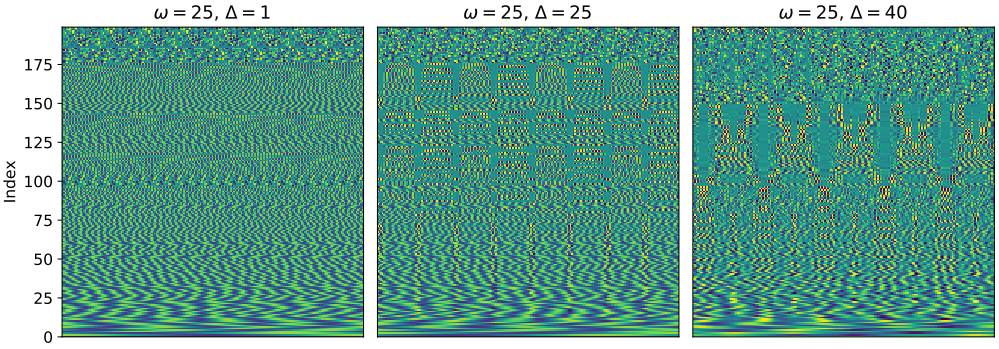

Figure 4: Heatmap of the eigenvectors $\{\boldsymbol{v}_j\}_{j\in[d]}$ ordered from highest associated eigenvalue (bottom) to lowest (top), for $d = 200, q = 30, r = 1, \omega = 25$, and $\Delta = 1$ (left), $\Delta = \omega = 25$ (middle) and $\Delta = 40$ (right).

From these observations, we expect $H^{\mathsf{CK}}_{\omega,\Delta}$ to have the same statistical properties as H^{CK}_{ω} when learning low-frequency functions. In Figure 5, we plot the test error of kernel ridge regression for fitting cyclic q-local polynomials (see Section A.7) on the hypercube of dimension $d = 30$. We report the test error of one realization, against the sample size n, and choose regularization $\lambda = 10^{-6}$ and noise $\sigma_\varepsilon = 0$. We compare kernels with and without downsampling. On the left, we consider $q = 10$ and $\omega = \Delta = 5$, and compare the test error with H^{CK}_{ω} (continous line) and with $H^{\mathsf{CK}}_{\omega,\Delta}$ (dashed line) when learning degree 2, 3 and 4 polynomials. On the right, we fix the target function to be the cubic local cyclic polynomial and consider the test error of learning with $H^{\mathsf{CK}}_{\omega,\Delta}$ for $q = 10$, $\omega = 10$, and $\Delta \in \{1, 3, 6, 10\}$. As expected, we observe in both simulations that the test error is almost identical between the kernels with and without downsampling, when learning cyclic invariant functions.

In Section C.1, we further check that downsampling with $\Delta > \omega$ does not improve the high-dimensional predictions for the test error of KRR.

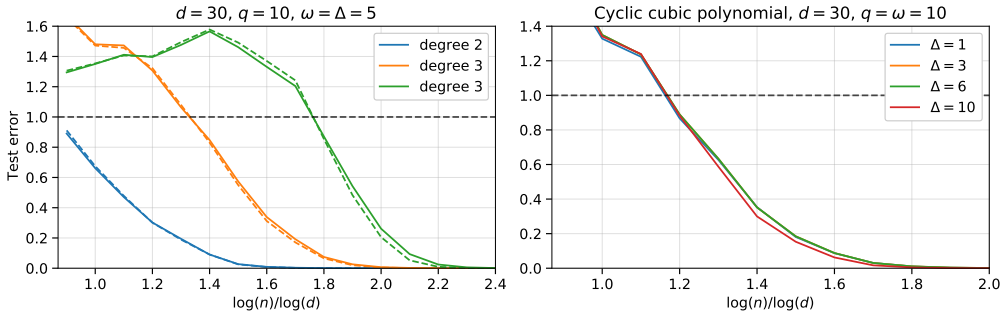

Figure 5: Test error of kernel ridge regression with and without downsampling. We report the test error of one realization, against the sample size n. On the left, we consider a unique architecture $q = 10$ and $\omega = \Delta = 5$, and compare H^{CK}_{ω} (continuous line) versus $H^{\mathsf{CK}}_{\omega,\Delta}$ (dashed line) when learning cyclic q-local polynomials of degree 2, 3 and 4. On the right, we consider a unique cyclic q-local polynomial of degree 3 for fixed $q = 10$, $\omega = 10$ and $\Delta \in \{1, 3, 6, 10\}$.

A.5 Multilayer convolutional kernels

For completeness, we briefly discuss here some intuitions of multilayer convolutional kernels. The benefit of depth in convolutional kernels has been investigated in [6, 16, 40, 43]. In particular, [6] observed that the top layer operation of a two-layers convolutional kernel can be replaced by a low-degree polynomial without a performance change.

As an example, we will consider a two layers convolutional kernel with patch and local average pooling sizes (q_1, ω_1) on the first layer and (q_2, ω_2) on the second layer. We consider a general inner-product kernel for the first layer:

$$h_1\big(\langle \boldsymbol{u}, \boldsymbol{v}\rangle / q_1\big) = \langle \psi(\boldsymbol{u}), \psi(\boldsymbol{v})\rangle \,, \tag{25}$$

where the feature map is given explicitly $\psi(\boldsymbol{u}) = \{\xi_{q_1,|S|} Y_S(\boldsymbol{u})\}_{S\subseteq[q_1]} \in \mathbb{R}^{2^{q_1}}$. Following the work [6], we consider a degree-2 polynomial kernel on the second layer, i.e., $h_2(\langle \phi, \phi'\rangle) = \langle \phi, \phi'\rangle^2$.

Let us decompose this two-layers convolutional kernel in the Fourier basis. Let $\Psi(\boldsymbol{x}) = \{\Psi_k(\boldsymbol{x})\}_{k\in[d]}$ be the output of the first layer, with

$$\Psi_k(\boldsymbol{x}) = \sum_{s\in[\omega_1]} \psi(\boldsymbol{x}_{(k+s)}) = \Big\{\xi_{q_1,|S|} \sum_{s\in[\omega_1]} Y_{k+s+S}(\boldsymbol{x})\Big\}_{S\subseteq[q_1]} \in \mathbb{R}^{2^{q_1}} \,. \tag{26}$$

Then denoting $\Psi_{(k)}(\boldsymbol{x}) = (\Psi_{k+1}(\boldsymbol{x}), \dots, \Psi_{k+q_2}(\boldsymbol{x}))$, the two-layers convolutional kernel is given by

$$\begin{aligned}
&H^{\mathsf{2CK}}_{\omega_1,\omega_2}(\boldsymbol{x},\boldsymbol{y}) \\
&= \sum_{k\in[d]} \sum_{s,s'\in[\omega_2]} \langle \Psi_{(k+s)}(\boldsymbol{x}), \Psi_{(k+s')}(\boldsymbol{x})\rangle^2 \\
&= \sum_{k\in[d]} \sum_{s,s'\in[\omega_2]} \sum_{u,u'\in[q_2]} \sum_{t,t',r,r'\in[\omega_1]} \\
&\qquad \langle \psi(\boldsymbol{x}_{(k+s+u+t)}) \otimes \psi(\boldsymbol{x}_{(k+s+u'+r)}), \psi(\boldsymbol{y}_{(k+s'+u+t')}) \otimes \psi(\boldsymbol{y}_{(k+s'+u'+r')})\rangle \,.
\end{aligned} \tag{27}$$

We believe that techniques contained in this paper can be used to study kernels of the type (**??**) by a careful combinatorial argument and a 2-dimensional Fourier transform on the second layer (see [6]). We leave this problem to future work. Here we only comment on the structure of $H^{\mathsf{2CK}}_{\omega_1,\omega_2}$:

1. Including a second convolutional layer allows interactions between patches. The associated RKHS, which we will denote $\mathcal{H}^{\mathsf{2CK}}$, contains all the homogeneous polynomials Y_S with $S = S_1 \cup S_2$ with S_1, S_2 contained on segments of size q_1, with the two segments separated by at most $q_2 + \omega_2 - 2$. In words, the RKHS contains interaction between patches $\boldsymbol{x}_{(k)}$ and $\boldsymbol{x}_{(k')}$ that are within some distance.
2. The eigenvalue associated to a degree-k homogeneous polynomials is still of order q^{-k} in high-dimension. To learn functions restricted to $L^2(\mathscr{Q}^2, \mathrm{Loc}_q)$, it is statistically more efficient to use H^{CK} (smaller degeneracy of eigenvalues). However H^{2CK} will fit a richer class of functions with two-patch interactions, while still not being plagued by dimensionality: $\dim(\mathcal{H}^{\mathsf{2CK}}) \le q_2 d 2^{2q_1}$. Hence we still expect H^{2CK} to be much more statistically efficient than a standard inner-product kernel.
3. Local pooling on the two layers plays different roles: pooling on the first layer encourages the interactions to not depend strongly on the relative positions of the patches, while pooling on the second layer penalizes functions that depend on the global position of these interactions.

For more layers and higher degree kernels, one obtain hierarchical interactions of higher-order, with multi-scale absolute and relative local invariances brought by pooling layers.

A.6 Proofs diagonalization of convolutional kernels

In this section, we prove the diagonalization of the kernels H^{CK}, H^{CK}_{ω} and $H^{\mathsf{CK}}_{\omega,\Delta}$ introduced in Propositions 1, 2 and 4 respectively.

Recall that we can associate to a kernel function $H : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined on a probability space $(\mathcal{X}, \tau)$ (assume $x \mapsto H(\boldsymbol{x}, \boldsymbol{x})$ square integrable), the integral operator $\mathbb{H} : L^2(\mathcal{X}, \tau) \to L^2(\mathcal{X}, \tau)$

$$\mathbb{H} f(\boldsymbol{x}) = \int_{\mathcal{X}} H(\boldsymbol{x}, \boldsymbol{x}') f(\boldsymbol{x}') \tau(\mathrm{d}\boldsymbol{x}') \,. \tag{28}$$

By the spectral theorem of compact operators, there exists an orthonormal basis $(\psi_j)_{j\ge1}$ of $L^2(\mathcal{X}, \tau)$ and eigenvalues $(\lambda_j)_{j\ge1}$, with nonincreasing values $\lambda_1 \ge \lambda_2 \ge \dots \ge 0$ and $\sum_{j\ge1} \lambda_j < \infty$, such that

$$\mathbb{H} = \sum_{j=1}^{\infty} \lambda_j \psi_j \psi_j^* , \qquad H(\boldsymbol{x}, \boldsymbol{x}') = \sum_{j=1}^{\infty} \lambda_j \psi_j(\boldsymbol{x}) \psi_j(\boldsymbol{x}') \,.$$

We first prove the diagonalization of $H^{\mathsf{CK}}_{\omega,\Delta}$ in Proposition 4. The case of H^{CK}_{ω} and H^{CK} then follows by setting $\Delta = 1$, and $\Delta = \omega = 1$ respectively.

Proof of Proposition 4. Consider the inner-product kernel function $h : \mathbb{R} \to \mathbb{R}$ defined on the hypercube $\mathscr{Q}^q$. By rotational symmetry (see Section 2.1 and Appendix D), h admits the following diagonalization: for any $\boldsymbol{u}, \boldsymbol{v} \in \mathscr{Q}^q$,

$$h\left(\langle \boldsymbol{u}, \boldsymbol{v}\rangle / q\right) = \sum_{\ell=0}^{q} \xi_{q,\ell} \sum_{S\subseteq[q], |S|=\ell} Y_S(\boldsymbol{u}) Y_S(\boldsymbol{v}) \,, \tag{29}$$

where $(Y_S)_{S\subseteq[q]}$ is the Fourier basis on $\mathscr{Q}^q$, and $\xi_{d,\ell}(h)$ is the ℓ-th Gegenbauer coefficient of h in dimension q (see Sections 2.1 or D for background).

Recall that we defined $\mathcal{S}_\ell = \{S \subseteq [q] : |S| = \ell\}$, the equivalence relation $S \sim S'$ if S' is a translated subset of S in $[q]$ (without cyclic convention), and $\mathcal{C}_\ell$ the quotient set of $\mathcal{A}_\ell$ by $\sim$. For each equivalence class $\overline{S} \in \mathcal{C}_\ell$, consider S the unique subset in $\overline{S}$ that contains 1. Then the equivalence class $\overline{S}$ contains the subsets $u + S = \{u + k : k \in S\} \subseteq [q]$ with $u = 0, \dots, q - \gamma(S)$. By a slight abuse of notations, we will identify $\overline{S}$ and this subset S. Below we will denote $u + S$ the translated subset with cyclic convention on $[d]$ (e.g., $2 + \{1, 3, d-1\} = \{3, 5, 1\}$).

Using Eq. (29) and that $Y_S(\boldsymbol{x}_{(k)}) = Y_{k+S}(\boldsymbol{x})$, we have the following decomposition of $H^{\mathsf{CK}}_{\omega,\Delta}$ in the Fourier basis

$$\begin{aligned} &H^{\mathsf{CK}}_{\omega,\Delta}(\boldsymbol{x},\boldsymbol{y}) \\ &= \frac{\Delta}{\omega} \sum_{k\in[d/\Delta]} \sum_{s,s'\in[\omega]} h\left(\langle \boldsymbol{x}_{(k\Delta+s)}, \boldsymbol{y}_{(k\Delta+s')}\rangle/q\right) \\ &= d\omega\xi_{q,0} + \sum_{\ell=1}^{q} \xi_{q,\ell} \sum_{S\in\mathcal{C}_\ell} \left\{ \frac{\Delta}{\omega} \sum_{(k,s,s',t)\in\mathcal{I}_{\omega,\Delta,\gamma(S)}} Y_{k\Delta+s+t+S}(\boldsymbol{x}) Y_{k\Delta+s'+t+S}(\boldsymbol{y}) \right\}, \end{aligned} \tag{30}$$

where we recall the definition of the set of indices

$$\mathcal{I}_{\omega,\Delta,\gamma(S)} = \left\{(k,s,s',t) : k \in [d/\Delta], s, s' \in [\omega], 0 \le t \le q - \gamma(S)\right\}. \tag{31}$$

Note that the diagonalization of the kernel H can be obtained by computing the matrix $\boldsymbol{M} = (M_{SS'})_{S,S'\subseteq[d]} \in \mathbb{R}^{2^d\times 2^d}$ with $M = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[Y_S(\boldsymbol{x})H(\boldsymbol{x},\boldsymbol{y})Y_{S'}(\boldsymbol{y})]$: if λ_j and $\boldsymbol{v}_j \in \mathbb{R}^{2^d}$ are the eigenvalues and eigenvectors of $\boldsymbol{M}$, then λ_j and $\psi_j(\boldsymbol{x}) = \sum_{S\subseteq[d]} v_{j,S} Y_S(\boldsymbol{x})$ are the eigenvalues and eigenvectors of H.

From Eq. (30), we see 1) the basis functions Y_S with $\gamma(S) > q$ (subset S not contained in a segment of size q) are in the null space of $H^{\mathsf{CK}}_{\omega,\Delta}$, 2) for $S, S' \subseteq [d]$ with S and S' not translations of each other, then $\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[Y_S(\boldsymbol{x})H^{\mathsf{CK}}_{\omega,\Delta}(\boldsymbol{x},\boldsymbol{y})Y_{S'}(\boldsymbol{y})] = 0$, and Y_S and $Y_{S'}$ are contained in orthogonal eigenspaces. We deduce that it is sufficient to diagonalize $H^{\mathsf{CK}}_{\omega,\Delta}$ on each of the (orthogonal) subspaces $V_S := \mathsf{span}\{Y_{k+S} : k \in [d]\}$ for $0 \le \ell \le q$ and $S \in \mathcal{C}_\ell$.

For each $S \in \mathcal{C}_\ell$, define $\boldsymbol{M}^{\gamma(S)} \in \mathbb{R}^{d\times d}$ the matrix with entries $M^{\gamma(S)}_{ij} = \frac{1}{r(S)}\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[Y_{i+S}(\boldsymbol{x})H^{\mathsf{CK}}_{\omega,\Delta}(\boldsymbol{x},\boldsymbol{y})Y_{j+S}(\boldsymbol{y})]$. From Eq. (30), we get

$$M^{\gamma(S)}_{ij} = \frac{\Delta}{\omega r(S)} \left|\left\{(k,s,s',t) \in \mathcal{I}_{\omega,\Delta,\gamma(S)} : k\Delta + s + t \equiv i[d], k\Delta + s' + t \equiv j[d]\right\}\right|, \tag{32}$$

which concludes the proof of Proposition 4. □

We can now prove Propositions 1 and 2 by taking $\omega = \Delta = 1$ and $\Delta = 1$ respectively.

Proof of Proposition 1. Set $\Delta = \omega = 1$ in Proposition 4. We get

$$\begin{aligned} \boldsymbol{M}^{\gamma(S)}_{ij} &= \frac{1}{r(S)} \left|\left\{(k,t) : k \in [d], 0 \le t \le q - \gamma(S), k + 1 + t \equiv i[d], k + 1 + t \equiv j[d]\right\}\right| \\ &= \delta_{ij}\,. \end{aligned}$$

In this case, $\boldsymbol{M}^{\gamma(S)}$ is simply equal to identity, which concludes the proof. □

Proof of Proposition 2. Set $\Delta = 1$ in Proposition 4. We get

$$\begin{aligned} \boldsymbol{M}^{\gamma(S)}_{ij} &= \frac{1}{\omega r(S)} \left|\left\{(k,s,s',t) \in \mathcal{I}_{\omega,\Delta,\gamma(S)} : k + s + t \equiv i[d], k + s' + t \equiv j[d]\right\}\right| \\ &= \left(1 - \frac{d(i,j)}{\omega}\right)_+, \end{aligned}$$

where $d(i,j)$ is the distance between i and j on the torus $[d]$ (i.e., if $i > j$, $d(i,j) = \min(i-j, d+j-i)$). Hence, $\boldsymbol{M}^{\gamma(S)}$ is a circulant matrix independent of $\gamma(S)$, which has well known explicit formula for eigenvalues and eigenvectors (see for example Remark 1). $\square$

A.7 Additional numerical simulations

Here, we consider a numerical experiment similar to Figure 1. We consider $\boldsymbol{x} \sim \mathrm{Unif}(\mathscr{Q}^d)$ with $d = 30$ and consider three cyclic invariant target functions:

$$f_2(\boldsymbol{x}) = \frac{1}{\sqrt{d}} \sum_{i \in [d]} x_i x_{i+1} \,, \qquad f_3(\boldsymbol{x}) = \frac{1}{\sqrt{d}} \sum_{i \in [d]} x_i x_{i+1} x_{i+2} \,,$$

$$f_4(\boldsymbol{x}) = \frac{1}{\sqrt{d}} \sum_{i \in [d]} x_i x_{i+1} x_{i+2} x_{i+3} \,.$$

We consider a higher order polynomial kernel $h(x) = \sum_{k \in [7]} 0.2 \cdot x^k$ than in Figure 1, which should lead to higher self-induced regularization. We consider the same kernels as before, with $q = 10$ and $\omega = 5$.

In Figure 6, we report the test errors of fitting f_2 (top), f_3 (middle) and f_4 (bottom) using kernel ridge regression with the 5 kernels of interests in the main text. We choose a small regularization parameter $\lambda = 10^{-6}$, and the noise level $\sigma_\varepsilon = 0$. The curves are averaged over 5 independent instances and the error bar stands for the standard deviation of these instances. The results again match with our overall theoretical predictions. We report the predicted thresholds for the three functions:

1. For f_2 target: $q < d < dq/\omega < dq < d^2$ for $H^{\mathsf{CK}}_{\mathsf{GP}} < H^{\mathsf{FC}}_{\mathsf{GP}} < H^{\mathsf{CK}}_{\omega} < H^{\mathsf{CK}} < H^{\mathsf{FC}}$.
2. For f_3 target: $q^2 < dq^2/\omega < d^2 < dq^2 < d^3$ for $H^{\mathsf{CK}}_{\mathsf{GP}} < H^{\mathsf{CK}}_{\omega} < H^{\mathsf{CK}} < H^{\mathsf{FC}}_{\mathsf{GP}} < H^{\mathsf{FC}}$.
3. For f_4 target: $q^3 < dq^3/\omega < d^3 < dq^3 < d^4$ for $H^{\mathsf{CK}}_{\mathsf{GP}} < H^{\mathsf{CK}}_{\omega} < H^{\mathsf{CK}} < H^{\mathsf{FC}}_{\mathsf{GP}} < H^{\mathsf{FC}}$.

We see that the kernels, especially for f_4, perform much better than their theoretical high-dimension predictions: this can be explained by the low-dimensionality of the experiment where $q = 10$.

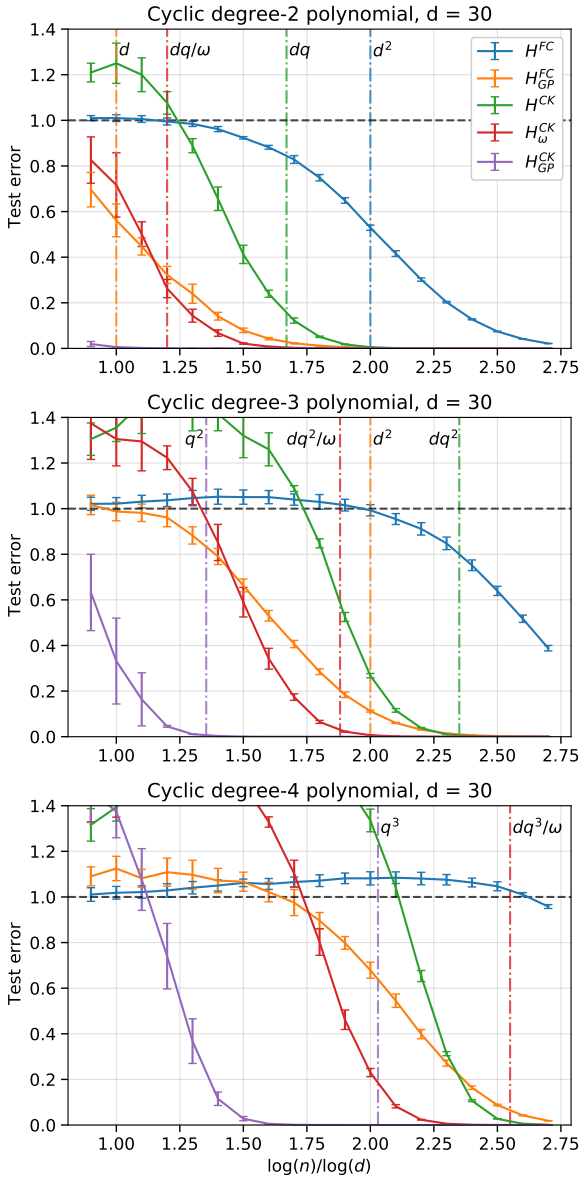

Figure 6: Learning cyclic polynomials of degree 2 (top), 3 (middle) and 4 (bottom) over the hypercube $d = 30$, using KRR with H^{FC} (FC), $H^{\mathsf{FC}}_{\mathsf{GP}}$ (FC-GP), H^{CK} (CK), H^{CK}_{ω} (CK-LP) and $H^{\mathsf{CK}}_{\mathsf{GP}}$ (CK-GP), regularization parameter $\lambda = 0^+$ and $h(x) = \sum_{k\in[7]} 0.2 \cdot x^k$. We report the average and the standard deviation of the test error over 5 realizations, against the sample size n.

B Generalization error of kernel methods in fixed dimension

B.1 Bound on kernel methods using Rademacher complexities

We first consider the case of a Lipschitz bounded loss and uniform convergence, and make a few simple remarks on the connection between generalization error and eigendecomposition in kernel methods.

Consider i.i.d data $(\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathbb{R}$ with $(\boldsymbol{x}, y) \sim P$ and a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ that we take 1-Lipschitz w.r.t second argument and bounded by 1. The goal is to minimize the expected loss $L(\hat{f}) = \mathbb{E}_{\boldsymbol{y},\boldsymbol{x}}\{\ell(y, \hat{f}(\boldsymbol{x}))\}$. Take a RKHS $\mathcal{H}$ with kernel function $H : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and consider following constrained empirical risk minimizer:

$$\hat{f}_B = \underset{\|f\|_{\mathcal{H}} \le B}{\arg\min} \left\{ \sum_{i=1}^{n} \ell(y_i, f(\boldsymbol{x}_i)) \right\} . \tag{33}$$

The generalization error of $\hat{f}_B$ has the following standard bound on the Rademacher complexity of the kernel class $\{f : \|f\|_{\mathcal{H}} \le B\}$ [9, 44]: with probability $1 - \delta$,

$$L(\hat{f}_B) - \min_{\|f\|_{\mathcal{H}} \le B} L(f) \le \frac{8B}{\sqrt{n}} \sqrt{\mathbb{E}_{\boldsymbol{x}}\{H(\boldsymbol{x}, \boldsymbol{x})\}} + \sqrt{\frac{2 \log \frac{2}{\delta}}{n}} . \tag{34}$$

Note that instead of a constraint on the norm in Eq. (33), one might find more convenient to use a penalty. In that case, there exists an equivalent to the bound (34) [3, 47], but we focus here on the constrained formulation for simplicity.

From the bound (34), we see that the generalization error depends crucially on the choice of B. For simplicity, let us forget about the approximation error and take $\|f_\star\|_{\mathcal{H}} \le B$ where $f_\star = \mathbb{E}\{y|\boldsymbol{x}\}$. Recall that for a kernel H with eigenvalues $\{\lambda_j\}_{j \ge 1}$ and eigenvectors $\{\psi_j\}_{j \ge 1}$, we have

$$\|f\|_{\mathcal{H}}^2 = \sum_{j \ge 1} \lambda_j^{-1} \langle \psi_j, f \rangle_{L^2(P)}^2 .$$

Consider $H^{\mathsf{CK}}_{\omega,\Delta}$ as in Eq. (8) and assume $\xi_{q,0} = 0$. From the normalization choice of the kernel (see Eq. (22)), we have

$$\mathbb{E}_{\boldsymbol{x}}\{H^{\mathsf{CK}}_{\omega,\Delta}(\boldsymbol{x}, \boldsymbol{x})\} = h(1).$$

Consider now for simplicity $\Delta = 1$. From the eigendecomposition in Proposition 2, the RKHS norm of $f \in L^2(\mathscr{Q}^d, \mathrm{Loc}_q)$ is given by

$$\|f\|_{\mathcal{H}}^2 = \sum_{\ell \in [q]} \sum_{j \in [d]} \sum_{S \in \mathcal{C}_\ell} \frac{\langle \psi_{j,S}, f \rangle_{L^2}^2}{\xi_{q,\ell} r(S) \kappa_j / d} .$$

Consider the case where $f \in L^2(\mathscr{Q}^d, \mathrm{Loc}_q)$ has a unique non-zero component in its discrete Fourier transform, i.e., $f(\boldsymbol{x}) = \frac{1}{\sqrt{d}} \sum_{k \in [d]} \rho_j^k g(\boldsymbol{x}_{(k)})$ with $\mathbb{E}\{g(\boldsymbol{x})\} = 0$ and $\rho_j = e^{2i\pi j/d}$ (see Section A.3). Note that, denoting $c_S = \langle Y_S, g \rangle_{L^2(\mathscr{Q}^q)}$:

$$f(\boldsymbol{x}) = \sum_{\ell=1}^{q} \sum_{S \in \mathcal{C}_\ell} \left(\sum_{u=0}^{r(S)-1} \rho_j^{-u} c_{u+S} \right) \psi_{j,S} .$$

Hence,

$$\|f\|_{\mathcal{H}}^2 = \sum_{\ell=1}^{q} \sum_{S \in \mathcal{C}_\ell} \frac{\langle \psi_{j,S}, f \rangle_{L^2}^2}{\xi_{q,\ell} r(S) \kappa_j / d} \le d \sum_{\ell=1}^{q} \sum_{S \in \mathcal{C}_\ell} \sum_{u=0}^{r(S)-1} \frac{c_{u+S}^2}{\xi_{q,\ell} r(S)} \le \frac{d \|g\|_h^2}{\kappa_j} ,$$

where $\|g\|_h^2$ is the RKHS norm associated to the inner-product kernel $h : \mathbb{R} \to \mathbb{R}$ in $\mathscr{Q}^q$, i.e., $\|g\|_h^2 = \sum_{S \subseteq [q]} \frac{c_S}{\xi_{q,|S|}}$. From the bound (34), we deduce the first generalization bound using a convolutional kernel: with probability at least $1 - \delta$,

$$L(\hat{f}_B) - \min_{\|f\|_{\mathcal{H}} \le B} L(f) \le 8 \left(\frac{d \|g\|_h^2 h(1)}{n \kappa_j} \right)^{1/2} + \sqrt{\frac{2 \log \frac{2}{\delta}}{n}} .$$

We make the following two remarks on this bound:

1. It depends on $\|g\|_h$, which is a RKHS norm on $\mathscr{Q}^q$ instead of $\mathscr{Q}^d$, which has potentially much lower dimension and contain less smooth function for balls of same radius.
2. There is a factor κ_j gain in sample complexity when learning functions that have j-th frequency with $\kappa_j > 1$. In particular, for $j = d$ (cyclic invariant functions), $\kappa_j = \omega$, and we need ω less samples to get the same (upper) bound on the generalization error. On the contrary, when $\kappa_j < 1$, i.e., high-frequency oscillatory functions, the generalization bound becomes worse.

B.2 Generalization error of KRR in the classical regime

We consider here the regression setting which allows for finer results. Several works have considered bounding the generalization error of kernel ridge regression (KRR) [12, 29], [47, Theorem 13.17]. In this section, we consider the following fully-explicit upper bound from [3].

Consider i.i.d data $(\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathbb{R}$ with $\boldsymbol{x}_i \sim P$, and $y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i$. Assume the noise $\mathbb{E}[\varepsilon_i|\boldsymbol{x}_i] = 0$ and $\mathbb{E}[\varepsilon_i^2|\boldsymbol{x}_i] \leq \sigma_\varepsilon^2$, and denote $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)$.

Let $\mathcal{H}$ be a RKHS with reproducing kernel $H : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The KRR solution with regularization parameter $\lambda \geq 0$ is given by

$$\hat{f}_\lambda = \arg\min_{f \in \mathcal{H}} \left\{ \sum_{i=1}^n (y_i - f(\boldsymbol{x}_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\} ,$$

which has the following analytical formula:

$$\hat{f}_\lambda(\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{x})(\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{y} ,$$

where $\boldsymbol{H} = (H(\boldsymbol{x}_i, \boldsymbol{x}_j))_{ij \in [n]}$ is the empirical kernel matrix, $\boldsymbol{h}(\boldsymbol{x}) = [H(\boldsymbol{x}, \boldsymbol{x}_1), \dots, H(\boldsymbol{x}, \boldsymbol{x}_n)]$ and $\boldsymbol{y} = (y_1, \dots, y_n)$. The risk is taken to be the test error with squared error loss

$$R(f_\star, \hat{f}_\lambda) = \mathbb{E}_{\boldsymbol{x}} \left\{ \left(f_\star(\boldsymbol{x}) - \hat{f}_\lambda(\boldsymbol{x}) \right)^2 \right\} . \tag{35}$$

Below, we give an upper bound on the expected risk over the noise $\boldsymbol{\varepsilon}$ in the training data, i.e., $\mathbb{E}_{\boldsymbol{\varepsilon}}\{R(f_\star, \hat{f}_\lambda)\}$ (it is also possible to give high probability bounds by concentration arguments, but we restrict ourselves to bounding the expected risk).

Theorem 6. *[3, Theorem 7.2] Assume $H(\boldsymbol{x}, \boldsymbol{x}) \leq R^2$ almost surely and let the regularization parameter $\lambda \leq R^2$. If $n \geq \frac{5R^2}{\lambda} \left(1 + \log \frac{R^2}{\lambda}\right)$, then*

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\{R(f_\star, \hat{f}_\lambda)\} \leq 16 \frac{\sigma_\varepsilon^2}{n} \mathcal{N}(H, \lambda) + 16 \inf_{f \in \mathcal{H}} \left\{ \|f - f_\star\|_{L^2}^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\} + \frac{24}{n^2} \|f_\star\|_{L^\infty}^2 , \tag{36}$$

where $\mathcal{N}(H, \lambda) = \mathrm{Tr}[(\mathbb{H} + \lambda \mathbf{I})^{-1} \mathbb{H}]$.

Let us comment on the upper-bound in Eq. (36). The first term corresponds to an upper bound on the variance: $\mathcal{N}(H, \lambda)$ is sometimes called the *degrees of freedom* or the *effective dimension* of the kernel H. The second term bounds the bias term and corresponds to an approximation error. In particular, for any $r > 0$,

$$\inf_{f \in \mathcal{H}} \left\{ \|f - f_\star\|_{L^2}^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\} \leq \lambda^r \|\mathbb{H}^{-r/2} f_\star\|_{L^2}^2 , \tag{37}$$

where we recall that $\mathbb{H}$ is the integral operator associated to H (see Eq. (28)). The third term can be removed by a more intricate analysis.

From the above discussion, it is natural to consider the following two assumptions on H and $f_\star$, that are standard in the kernel literature:

(B1) *Capacity condition:* $\mathcal{N}(H, \lambda) \leq C_H \lambda^{-1/\alpha}$ with $\alpha > 1$.

(B2) *Source condition:* there exists $\beta > 0$ such that $\|\mathbb{H}^{-\beta/2} f_\star\|_{L^2}^2 =: B_{f_\star}^2 < \infty$.

Intuitively, the capacity condition (B1) characterizes the size of the RKHS: for increasing α, the RKHS contains less and less functions. It is verified when the eigenvalues λ_j's of H decay at the rate $j^{-\alpha}$. For example, taking the Matern kernel of order $s > d/2$, whose RKHS is the Sobolev space of order s (i.e., functions with bounded s-order derivatives), we have $\alpha = 2s/d$ (e.g., see [26]). The source condition (B2) characterizes the regularity of the target function (the 'source') with respect to the kernel: $\beta = 1$ is equivalent to $f_\star \in \mathcal{H}$, while $\beta > 1$ corresponds to $f_\star$ more smooth (and $\beta < 1$ less smooth $f_\star$).

Assuming (B1) and (B2) in Theorem 6, we get the bound

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\varepsilon}}\{R(f_\star, \hat{f}_\lambda)\} &\leq 16 C_H \frac{\sigma_\varepsilon^2}{n}\lambda^{-1/\alpha} + 16 B_{f_\star}^2 \lambda^\beta + \frac{24}{n^2}\|f_\star\|_{L^\infty}^2 \\ &= 32\sigma_\varepsilon^2 B_{f_\star}^{\frac{2}{\alpha\beta+1}} \left(\frac{C_H}{n}\right)^{\frac{\alpha\beta}{\alpha\beta+1}} + \frac{24}{n^2}\|f_\star\|_{L^\infty}^2 \,,\end{aligned} \tag{38}$$

where in the second line, we balanced the two terms by taking $\lambda_* := \left(\frac{C_H \sigma_\varepsilon^2}{B_{f_\star}^2 n}\right)^{\frac{\alpha}{\alpha\beta+1}}$. Note that in order to use Theorem 6, we need further to constrain $n \geq \frac{5R^2}{\lambda}\left(1 + \log\frac{R^2}{\lambda}\right)$. For simplicity, we will choose $r > \frac{\alpha-1}{\alpha}$, so that this condition is verified for n sufficiently large.

Remark 2. The rate in n in Eq. (38) is minmax optimal over all functions that verify assumptions (A1) and (A2) [12]. However, for large d, the RKHS is composed of very smooth functions (e.g., Sobolev spaces of order s are RKHS if and only if $s > d/2$, i.e., if the order of the bounded derivatives grows with the dimension d) and β will be small, such that $\beta\alpha \approx \kappa/d$ for functions with bounded derivatives up to order κ. In that case, the risk decreases at the rate $n^{-O(\frac{\kappa}{d})}$: KRR suffers from the curse of dimensionality when κ does not scale with d. As a consequence, the bound (38) is vacuous when n does not scale exponentially in d, which led several groups to derive finer bounds on KRR in the high dimensional regime (see Section C).

Let us now apply Theorem 6 and Eq. (38) to our convolutional kernels to show Theorems 1 and 4.

Proof of Theorem 1. First notice that $H^{\mathsf{CK}}(\boldsymbol{x}, \boldsymbol{x}) = h(1) =: R^2$ and we can therefore apply Theorem 6. The effective dimension of H^{CK} is bounded by

$$\begin{aligned}\mathcal{N}(\mathbb{H}^{\mathsf{CK}}, \lambda) &= \frac{\xi_{q,0}}{\xi_{q,0} + \lambda} + \sum_{\ell=1}^{q} \sum_{S \in \mathcal{E}_\ell} \frac{\xi_{q,\ell} r(S)/d}{\xi_{q,\ell} r(S)/d + \lambda} \\ &\leq \frac{d\xi_{q,0}}{\xi_{q,0} + d\cdot\lambda} + \sum_{\ell=0}^{q} \frac{\xi_{q,\ell}}{\xi_{q,\ell} + d\cdot\lambda} \sum_{S\in\mathcal{E}_\ell} r(S) \\ &= d\sum_{\ell=0}^{q} B(\mathscr{Q}^q, \ell) \frac{\xi_{q,\ell}}{\xi_{q,\ell} + d\cdot\lambda} = d\mathcal{N}(h, d\cdot\lambda)\,,\end{aligned}$$

where we used that $r(S) \geq 1$ in the second line and $\mathcal{N}(h, \lambda)$ is the effective dimension of the inner-product kernel h on $\mathscr{Q}^q$. We deduce from (A1) that $\mathcal{N}(H^{\mathsf{CK}}, \lambda) \leq C_h d^{1-1/\alpha}\lambda^{-1/\alpha}$. Furthermore, from (A2) and the assumption that $\mathbb{E}\{g_k(\boldsymbol{x})\} = 0$, we have

$$\begin{aligned}\|(H^{\mathsf{CK}})^{-\beta/2} f_\star\|_{L^2}^2 &= d^\beta \sum_{\ell=1}^{q} \xi_{q,\ell}^{-\beta} \sum_{S\in\mathcal{C}_\ell} \sum_{k\in[d]} r(S)^{-\beta} \left(\sum_{u=0}^{r(S)-1} \langle g_{k-u}, Y_{u+S}\rangle_{L^2}\right)^2 \\ &\leq d^\beta \sum_{\ell=1}^{q} \xi_{q,\ell}^{-\beta} \sum_{S\in\mathcal{C}_\ell} \sum_{k\in[d]} r(S)^{1-\beta} \sum_{u=0}^{r(S)-1} \langle g_{k-u}, Y_{u+S}\rangle_{L^2}^2 \\ &\leq d^\beta q^{1-\beta} \sum_{k=1}^{d} \|h^{-\beta/2} g_k\|_{L^2}^2 \leq d^\beta q B^2\,.\end{aligned}$$

Injecting the two above bounds in Eq. (38), we deduce that there exists constants C_1, C_2, C_3 that only depends on the constants in (A1) and (A2), and $h(1), \sigma_\varepsilon^2$ (but independent of d), such that taking

$n \geq C_1 \max(\|f_\star\|_{L^\infty}^2, d)$ and $\lambda_* = \frac{C_2}{d}(d/n)^{\frac{\alpha}{\alpha\beta+1}}$, we get

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\{R(f_\star, \hat{f}_{\lambda_\star})\} \leq C_3 \left(\frac{d}{n}\right)^{\frac{\alpha\beta}{\alpha\beta+1}} .$$

□

Proof of Theorem 4. The proof is similar to the proof of Theorem 1. Notice that $H_\omega^{\mathsf{CK}}(\boldsymbol{x}, \boldsymbol{x}) \leq h(1)$, and that the effective dimension of H_ω^{CK} is bounded by

$$\begin{aligned}\mathcal{N}(H_\omega^{\mathsf{CK}}, \lambda) &= \sum_{j=1}^{d}\sum_{\ell=1}^{q}\sum_{S\in\mathcal{C}_\ell} \frac{\xi_{q,\ell} r(S)\kappa_j/d}{\xi_{q,\ell} r(S)\kappa_j/d + \lambda} \\ &\leq \sum_{j=1}^{d}\sum_{\ell=1}^{q}\sum_{S\in\mathcal{C}_\ell} r(S)\frac{\xi_{q,\ell}}{\xi_{q,\ell} + d\lambda/\kappa_j} = \sum_{j=1}^{d}\mathcal{N}(h, d\lambda/\kappa_j) \leq C_h d^{-1/\alpha}\lambda^{-1/\alpha}\sum_{j=1}^{d}\kappa_j^{1/\alpha},\end{aligned}$$

where we used condition (A1). Denoting $d_{\text{eff}} = \sum_{j=1}^{d}(\kappa_j/\omega)^{1/\alpha}$, the rest of the proof follows from the proof of Theorem 1 with d replaced by $d_{\text{eff}}\omega^{1/\alpha}$ and B^2 replaced by $\omega^\beta B^2$. □

Remark 3. Note that the requirement $\|(\mathbb{H}_\omega^{\mathsf{CK}}/\omega)^{-\beta/2} f_\star\|_{L^2} \leq B$ is to make the result comparable to the other theorems when we consider target functions with low-frequencies. For a cyclic invariant function, we get exactly $\|(\mathbb{H}_\omega^{\mathsf{CK}}/\omega)^{-\beta/2} f_\star\|_{L^2} = \|(\mathbb{H}^{\mathsf{CK}})^{-\beta/2} f_\star\|_{L^2}$.

C Generalization error of KRR in high dimension

In Section B.2, we considered upper bounds on the test error of KRR using the standard *capacity* and *source conditions*. However, these results suffer from several limitations:

1. They only provide an upper bound on the test error. While the decay rate with respect to n is minmax optimal (see [12]), this is not strong enough to show, for example, a statistical advantage of using local average pooling, which appears as a prefactor d_{eff}, and which would require a lower bound matching the upper bound within a constant factor.
2. As mentioned in Remark 2, the bound is of order $n^{-1/O(d)}$, except when the target function has smoothness order increasing with d. This bound is non-vacuous only if $n = \exp(O(d))$ which is impractical in modern image datasets where typically $d \geq 100$. This motivates a new type of question: given $n \asymp d^\alpha$, what is the prediction error achieved by KRR for a given function?
3. In order to achieve the bound Eq. (38), one need to carefully balance the bias and the variance terms by setting the regularization parameter. This is in contrast with modern practice which usually train until interpolation (which corresponds to setting $\lambda \to 0$).

Given the above limitations, several recent works have instead considered a high-dimensional setting where the number of samples scales with d, and derived asymptotic test errors, exact up to a vanishing additive error [23, 24, 38]. In addition to these works, several papers have derived general estimates for the test error using non-rigorous methods [11, 17, 29] that are believe to be correct in the high dimensional limit and which show great agreement with numerical experiments. The picture that emerges in this regime is much more precise than in the classical regime: KRR approximately acts as a *shrinkage operator* on the target function (not assumed to be in a particular space anymore), with shrinkage parameter that scales as a self-induced regularization parameter over the number of samples.

More precisely, [38] shows the following: considers a kernel $H_d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with eigenvalues $(\lambda_{d,j})_{j\geq 1}$ in nonincreasing order and $n \equiv n(d)$ the number of samples. Let $m \equiv m(d)$ be an integer such that $m \leq n^{1-\delta}$ and

$$\lambda_{d,m+1} \cdot n^{1+\delta} \leq \sum_{j=m+1}^{\infty} \lambda_{d,j} ,$$

for some $\delta > 0$. Then, assuming some additional conditions insuring that the kernel H_d is 'spread-out' and well behaved, the KRR solution

$$\hat{f}_\lambda = \arg\min_{f \in \mathcal{H}_d} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f(\boldsymbol{x}_i) \right)^2 + \frac{\lambda}{n} \|f\|_{\mathcal{H}_d}^2 \right\} , \tag{39}$$

is equal up to a vanishing additive L^2-error (as $d \to \infty$) to the following effective ridge regression estimator

$$\hat{f}^{\text{eff}}_{\lambda_{\text{eff}}} = \arg\min_{f \in \mathcal{H}_d} \left\{ \|f_\star - f\|_{L^2}^2 + \frac{\lambda_{\text{eff}}}{n} \|f\|_{\mathcal{H}_d}^2 \right\} , \tag{40}$$

where $\lambda_{\text{eff}} = \lambda + \sum_{j=m+1}^{\infty} \lambda_{d,j}$. The effective estimator (40) amounts to replacing the empirical risk in Eq. (39) by its population counterpart $\|f_\star - f\|_{L^2}^2 = \mathbb{E}_{\boldsymbol{x}}\{(f_\star(\boldsymbol{x}) - f(\boldsymbol{x}))^2\}$. In words, in high dimension, KRR with a finite number of samples is the same as KRR with infinite number of samples but with a larger ridge regularization.

The solution of Eq. (40) admits an explicit solution in terms of a *shrinkage operator* in the basis $(\psi_{d,j})_{j \geq 1}$ of eigenfunctions of H_d:

$$f_\star(\boldsymbol{x}) = \sum_{j=1}^{\infty} c_j \psi_{d,j}(\boldsymbol{x}) \qquad \mapsto \qquad \hat{f}^{\text{eff}}_{\lambda_{\text{eff}}} = \sum_{j=1}^{\infty} \frac{\lambda_{d,j}}{\lambda_{d,j} + \frac{\lambda_{\text{eff}}}{n}} \cdot c_j \cdot \psi_{d,j}(\boldsymbol{x}) \, . \tag{41}$$

Hence, KRR will fit better the target function along eigendirections associated to larger eigenvalues of H. If $\lambda_{d,j} \gg \lambda_{\text{eff}}/n$, KRR fits perfectly $f_\star$ along the eigendirection $\psi_{d,j}$, while if $\lambda_{d,j} \ll \lambda_{\text{eff}}/n$, KRR does not fit this eigendirection at all. This phenomena has been referred as the *spectral bias* and *task-kernel alignment* of kernel ridge regression in several works.

Finally, notice from Eq. (41) that the minimum test error is achieved for the regularization parameter $\lambda = 0$, which corresponds to the KRR estimator fitting perfectly the training data. In other words, the *interpolating solution is optimal for kernel ridge regression in high dimension.*

C.1 Generalization error of convolutional kernels in high dimension

Consider a sequence of integers $\{d(q)\}_{q \geq 1}$ which corresponds to a sequence of image spaces $\boldsymbol{x} \in \mathscr{Q}^d$ of increasing dimension, and assume $d(q)/2 \geq q \geq d(q)^\delta$ for some constant $\delta > 0$. For ease of notations, we will keep the dependency on q implicit, i.e., $d := d(q)$. Let $\{h_q\}_{q \geq 1}$ be a sequence of inner-product kernels $h_q : \mathbb{R} \to \mathbb{R}$.

Test error with one-layer convolutional kernel: we first consider a vanilla one-layer convolutional kernel H^{CK} as defined in Eq. (3). We will assume that the kernels $\{h_q\}_{q \geq 1}$ verify the following 'genericity' condition.

Assumption 1 (Generecity assumption on $\{h_q\}_{q \geq 1}$ at level $\mathsf{s} \in \mathbb{N}$)**.** *For $\{h_q\}_{q \geq 1}$ a sequence of inner-product kernels $h_q : \mathbb{R} \to \mathbb{R}$, we assume the following conditions to hold. There exists $\mathsf{s}' \geq 1/\delta + 2\mathsf{s} + 3$ where $\delta > 0$ verifies $q \geq d^\delta$ and a constant C such that $h_q(1) \leq C$, and*

$$\min_{k \leq \mathsf{s}-1} q^{\mathsf{s}-1-k} \xi_{q,k} B(q,k) = \Omega_d(1), \tag{42}$$

$$\min_{k \in \{\mathsf{s}, \mathsf{s}+1, \mathsf{s}'\}} \xi_{q,k} B(q,k) = \Omega_d(1), \tag{43}$$

$$\max_{k=0,\dots,\mathsf{s}'} q^{\mathsf{s}'-k+1} \xi_{q,q-k} B(q, q-k) = O_d(1). \tag{44}$$

Assumption 1 will be verified by standard kernels, e.g., the Gaussian kernel. We discuss this assumption in Section C.2 and present sufficient conditions on the activation function σ for its associated CNTK to verify Assumption 1.

Recall that we denoted $L^2(\mathscr{Q}^d, \mathrm{Loc}_q)$ the space of local functions, i.e., that can be decomposed as $f(\boldsymbol{x}) = \sum_{k \in [d]} f_k(\boldsymbol{x}_{(k)})$. Denote $h_{q,>\ell}$ the inner-product kernel h_q with its $(\ell+1)$-first Gegenbauer coefficients set to 0, i.e.,

$$h_{q,>\ell}(\langle \boldsymbol{u}, \boldsymbol{v} \rangle / q) = \sum_{k=\ell+1}^{q} \xi_{q,k} B(\mathscr{Q}^q; k) Q_k^{(q)}(\langle \boldsymbol{u}, \boldsymbol{v} \rangle) \, , \tag{45}$$

for any $\boldsymbol{u}, \boldsymbol{v} \in \mathscr{Q}^q$. The following result is a consequence of the general theorem on the generalization error of KRR in [38].

Theorem 7 (Test error of CK in high dimension)**.** *Let $\{f_d \in L^2(\mathscr{Q}^d, \mathrm{Loc}_q)\}_{q\geq 1}$ be a sequence of local functions. Let $(\boldsymbol{x}_i)_{i\in[n(d)]} \sim_{\text{i.i.d.}} \mathrm{Unif}(\mathscr{Q}^d)$ and $y_i = f_d(\boldsymbol{x}_i) + \varepsilon_i$ with $\varepsilon_i \sim_{\text{i.i.d.}} \mathsf{N}(0, \sigma_\varepsilon^2)$. Assume $d \cdot q^{\mathsf{s}-1+\delta} \leq n \leq d \cdot q^{\mathsf{s}-\delta}$ for some $\delta > 0$ and let $\{h_q\}_{q\geq 1}$ be a sequence of activation functions satisfying Assumption 1 at level s. Consider $\{H^{\mathsf{CK},d}\}_{q\geq 1}$ the sequence of convolutional kernels associated to $\{h_q\}_{q\geq 1}$ as defined in Eq.* (3)*. Then the following holds for the solution $\hat{f}_\lambda$ of KRR with kernels $\{H^{\mathsf{CK},d}\}_{q\geq 1}$.*

For any regularization parameter $\lambda \geq 0$, define the effective regularization $\lambda_{\text{eff}} := \lambda + h_{q,>\mathsf{s}}(1)$. Then for any $\eta > 0$, we have

$$\left\|\hat{f}_\lambda - \hat{f}^{\text{eff}}_{\lambda_{\text{eff}}}\right\|_{L^2}^2 = o_{d,\mathbb{P}}(1) \cdot (\|f_d\|_{L^{2+\eta}}^2 + \sigma_\varepsilon^2). \tag{46}$$

The proof of Theorem 7 is deferred to Section C.4.

Let us expound on the predictions of Theorem 7. First, recall that $\hat{f}^{\text{eff}}_{\lambda_{\text{eff}}}$ is given explicitly in Eq. (41) by a shrinkage operator with parameter λ_{eff}. From Assumption 1 and taking $\lambda = 0$, the shrinkage operator is of order 1

$$\lambda_{\text{eff}} = h_{q,>\mathsf{s}}(1) = \sum_{\ell=\mathsf{s}+1}^{q} \xi_{q,\ell} B(\mathscr{Q}^q; \ell) = \Theta_q(1)\,.$$

From the eigendecomposition of H^{CK} introduced in Proposition 1, KRR fits perfectly $f_\star$ along the eigendirection Y_S with $|S| = \ell$ if $n \cdot \xi_{d,\ell} r(S)/d \gg \lambda_{\text{eff}}$, while it does not fit this eigendirection at all if $n \cdot \xi_{d,\ell} r(S)/d \leq \lambda_{\text{eff}}$. Consider $n = d \cdot q^{\mathsf{s}-1+\alpha}$:

- KRR fits the eigendirections corresponding to the homogeneous polynomials of degree $\mathsf{s}-1$ and less, and of degree s for subsets S such that $\gamma(S) \ll q - q^{1-\alpha}$.
- KRR does not fit at all the eigendirections correpsonding to homogeneous polynomials of degree $\mathsf{s}+1$ and larger, and degree s for subsets S such that $\gamma(S) \gg q - q^{1-\alpha}$.

In words, for $d \cdot q^{\mathsf{s}-1} \ll n \ll d \cdot q^{\mathsf{s}}$, KRR fits at least a degree-$(\mathsf{s}-1)$ polynomial approximation to $f_\star$ and at most a degree-s polynomial approximation. As n increases from $d \cdot q^{\mathsf{s}-1}$ to $d \cdot q^{\mathsf{s}}$, KRR first fits degree-s homogeneous polynomials that have smaller diameter $\gamma(S)$ (i.e., 'more localized').

Test error of CK with global average pooling: we consider the kernel $H^{\mathsf{CK}}_{\mathsf{GP}}$ given by a convolutional layer followed by global average pooling:

$$H^{\mathsf{CK}}_{\mathsf{GP}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{d} \sum_{k,k'\in[d]} h\left(\langle \boldsymbol{x}_{(k)}, \boldsymbol{y}_{(k')}\rangle/q\right)\,, \tag{47}$$

In addition to the genericity condition, we will assume that the kernels $\{h_q\}_{q\geq 1}$ verify the following differentiability condition.

Assumption 2 (Differentiability assumption on $\{h_q\}_{q\geq 1}$ at level $\mathsf{s} \in \mathbb{N}$)**.** *For $\{h_q\}_{q\geq 1}$ a sequence of inner-product kernels $h_q : \mathbb{R} \to \mathbb{R}$, we assume the following conditions to hold. There exists $v \geq \max(2/\delta, \mathsf{s})$ where $\delta > 0$ verifies $q \geq d^\delta$ such that h_q is $(v+1)$-differentiable and for $k \leq v$,*

$$\sup_{\gamma\in[-1,1]} \left|h^{(v+1)}_{q,>v}(\gamma)\right| \leq O_q(1),$$

$$\left|h^{(k)}_{q,>v}(0)\right| \leq O_q(q^{-(v+1-k)/2}),$$

where we denoted $h_{q,>v}$ the truncated inner-product kernel h_q as in Eq. (45)*.*

Assumption 2 is used to extend the following theorem to non-polynomial kernel h_q (in particular, it is trivially verified for polynomial kernels by taking v larger than the degree of h_q). This assumption is difficult to check in practice, however we provide some examples where it holds in Appendix C.2.

Recall that we denoted $L^2(\mathscr{Q}^d, \mathrm{CycLoc}_q)$ the space of functions that are given by the convolution of a function $g : \mathbb{R}^q \to \mathbb{R}$ with the image $\boldsymbol{x} \in \mathscr{Q}^d$, i.e., $f(\boldsymbol{x}) = \sum_{k\in[d]} g(\boldsymbol{x}_{(k)})$.

Theorem 8 (Test error of CK with GP in high dimension)**.** *Let $\{f_d \in L^2(\mathscr{Q}^d, \mathrm{CycLoc}_q)\}_{q\geq 1}$ be a sequence of convolutional functions. Assume $q^{\mathsf{s}-1+\delta} \leq n \leq q^{\mathsf{s}-\delta}$ for some $\delta > 0$ and let $\{h_q\}_{q\geq 1}$ be a sequence of activation functions satisfying Assumptions 1 and 2 at level s. Consider $\{H_{\mathsf{GP}}^{\mathsf{CK},d}\}_{q\geq 1}$ the sequence of convolutional kernels with global pooling associated to $\{h_q\}_{q\geq 1}$ as defined in Eq.* (47)*. Then the solution $\hat{f}_\lambda$ of KRR with kernels $\{H_{\mathsf{GP}}^{\mathsf{CK},d}\}_{q\geq 1}$ verifies Eq.* (46) *with $\lambda_{\text{eff}} := \lambda + h_{q,>\mathsf{s}}(1)$.*

The proof of Theorem 8 is deferred to Section C.5.

The predictions of Theorem 8 are similar to the ones of Theorem 7 but with a factor d gain in statistical efficiency: this is due to the eigenvalues of $H_{\mathsf{GP}}^{\mathsf{CK}}$ being a factor d larger than for H^{CK}. Therefore, with global average pooling, for $q^{\mathsf{s}-1} \ll n \ll q^{\mathsf{s}}$, KRR fits at least a degree-$(\mathsf{s}-1)$ invariant polynomial approximation to $f_\star$ and at most a degree-s invariant polynomial approximation. As n increases from $q^{\mathsf{s}-1}$ to q^{s}, KRR first degree-s invariant homogeneous polynomials with increasing diameter $\gamma(S)$.

Test error of CK with local average pooling: In the case of local average pooling with $\omega < d$, the eigenvalues are harder to control. Indeed, we have mixing of the eigenvalues between polynomials of different degree: there exists $j, j' \in [d]$ such that $\xi_{q,\ell}\kappa_j \ll \xi_{q,\ell+1}\kappa_{j'}$. The eigenvalues are not ordered in increasing degree of their associated eigenfunctions anymore. While this case is potentially tractable with a more careful analysis, we instead introduce a simplified kernel which we believe qualitatively captures the statistical behavior of local average pooling.

Assume $q \leq \omega/2$ and ω is a divisor of d. Denote $\boldsymbol{x}^{(k\omega)} = (x_{k\omega+1}, \dots, x_{k\omega+\omega})$ the k-th segment of length ω in $[d]$ and $\boldsymbol{x}_{(i)}^{(k\omega)} = (x_{k\omega+i}, \dots, x_{k\omega+q+i})$ the patch of size q with cyclic convention in $\{k\omega+1, \dots, k\omega+\omega\}$. Consider the following convolutional kernel with 'non-overlapping' average pooling:

$$H_\omega^{\mathsf{CK},\mathsf{NO}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{\omega} \sum_{k\in[d/\omega]} \sum_{i,j\in[\omega]} h_q\big(\langle \boldsymbol{x}_{(i)}^{(k\omega)}, \boldsymbol{y}_{(j)}^{(k\omega)}\rangle/q\big)\,, \tag{48}$$

In words, $H_\omega^{\mathsf{CK},\mathsf{NO}}$ is the combination of d/ω non-overlapping convolutional kernels with global average pooling on images of size ω:

$$\begin{aligned} H_\omega^{\mathsf{CK},\mathsf{NO}} &= \sum_{k\in[d/\omega]} H_{\mathsf{GP}}^{\mathsf{CK}}\big(\boldsymbol{x}^{(k\omega)}, \boldsymbol{y}^{(k\omega)}\big) \\ &= \sum_{\ell=0}^{q} \xi_{q,\ell} \sum_{k\in[d/\omega]} \sum_{S\in\mathcal{C}_\ell} \psi_{k,S}(\boldsymbol{x})\psi_{k,S}(\boldsymbol{y})\,, \end{aligned} \tag{49}$$

where $\psi_{k,S}(\boldsymbol{x}) = \frac{1}{\sqrt{\omega}} \sum_{i\in[\omega]} Y_{i+S}(\boldsymbol{x}^{(k\omega)})$ where $i+S$ is the translated set with cyclic convention in $[\omega]$.

Denote $L^2(\mathscr{Q}^d, \mathrm{LocCycLoc}_q)$ the RKHS associated to $H_\omega^{\mathsf{CK},\mathsf{NO}}$, which contains functions that are locally convolutions on segments of size ω. For this simplified model, the proof of Theorem 8 can be easily adapted and we obtain the following result:

Corollary 1 (Test error of CK with NO pooling in high dimension)**.** *Let $\{f_d \in L^2(\mathscr{Q}^d, \mathrm{LocCycLoc}_q)\}_{q\geq 1}$ be a sequence of local convolutional functions. Assume $(d/\omega)\cdot q^{\mathsf{s}-1+\delta} \leq n \leq (d/\omega)\cdot q^{\mathsf{s}-\delta}$ for some $\delta > 0$ and let $\{h_q\}_{q\geq 1}$ be a sequence of activation functions satisfying Assumptions 1 and 2 at level s. Consider $\{H_\omega^{\mathsf{CK},\mathsf{NO},d}\}_{q\geq 1}$ the sequence of convolutional kernels with non-overlapping pooling associated to $\{h_q\}_{q\geq 1}$ as defined in Eq.* (48)*. Then the solution $\hat{f}_\lambda$ of KRR with kernels $\{H_\omega^{\mathsf{CK},\mathsf{NO},d}\}_{q\geq 1}$ verifies Eq.* (46) *with $\lambda_{\text{eff}} := \lambda + \frac{d}{\omega} h_{q,>\mathsf{s}}(1)$.*

Corollary 1 shows that $H_\omega^{\mathsf{CK},\mathsf{NO}}$ enjoys a factor ω gain in statistical efficiency compared to H^{CK}, due to a factor ω smaller effective ridge regularization. Therefore, with (non-overlapping) local average pooling, for $(d/\omega)\cdot q^{\mathsf{s}-1} \ll n \ll (d/\omega)\cdot q^{\mathsf{s}}$, KRR fits degree-$(\mathsf{s}-1)$ locally invariant polynomials and none of the polynomials of degree-$(\mathsf{s}+1)$ and larger. Heuristically, we see that this yields the same statistical efficiency than H^{CK} for $\omega = 1$ and $H_{\mathsf{GP}}^{\mathsf{CK}}$ for $\omega = d$, and interpolates between the two cases for $1 < \omega < d$.

Test error of convolutional kernels with downsampling: We consider adding a downsampling operation to the previous kernels. Let Δ be a constant and a divisor of d and ω and consider the following 'downsampled' kernels:

$$H_{\Delta}^{\mathsf{CK}}(\boldsymbol{x},\boldsymbol{y}) = \Delta \sum_{k\in[d/\Delta]} h\big(\langle \boldsymbol{x}_{(k\Delta)}, \boldsymbol{y}_{(k\Delta)}\rangle/q\big)\,, \tag{50}$$

$$H_{\mathsf{GP},\Delta}^{\mathsf{CK}}(\boldsymbol{x},\boldsymbol{y}) = \frac{\Delta}{d} \sum_{k,k'\in[d/\Delta]} h\big(\langle \boldsymbol{x}_{(k\Delta)}, \boldsymbol{y}_{(k'\Delta)}\rangle/q\big)\,, \tag{51}$$

$$H_{\omega,\Delta}^{\mathsf{CK},\mathsf{NO}}(\boldsymbol{x},\boldsymbol{y}) = \sum_{k\in[d/\omega]} H_{\mathsf{GP},\Delta}^{\mathsf{CK}}\big(\boldsymbol{x}^{(k\omega)}, \boldsymbol{y}^{(k\omega)}\big)\,. \tag{52}$$

We can easily adapt the proofs of Theorems 7 and 8, and Corollary 1 to these kernels. In particular, their conclusions do not change (for any constant Δ) and downsampling do not provide a statistical advantage.

C.2 Checking the assumptions

In this section, we discuss Assumptions 1 and 2 and present sufficient conditions for them to be verified.

Genericity assumption: Recall that the inner-product kernel $h_q : \mathbb{R} \to \mathbb{R}$ has the following eigendecomposition on $\mathscr{Q}^q$ as

$$h_q\big(\langle \boldsymbol{u},\boldsymbol{v}\rangle/q\big) = \sum_{\ell=0}^{q} \xi_{q,\ell} \sum_{S\subseteq[q],|S|=\ell} Y_S(\boldsymbol{u})Y_S(\boldsymbol{v})\,.$$

The genericity assumption amounts to: 1) A universality condition in Eqs. (42) and (43): if $P_k h(\langle \mathbf{1},\cdot\rangle/q) = 0$, then h does not learn degree-k homogeneous polynomials; 2) A constant order scaling of the self-induced regularization $h_{q,>\mathsf{s}}(1)$, from $h_q(1) \le C$ and Eq. (43) with s', i.e., $h_{q,>\mathsf{s}}(1) \le h_q(1) = O_q(1)$ and $h_{q,>\mathsf{s}}(1) \ge \xi_{q,\mathsf{s}'} B(q,\mathsf{s}') = \Omega_q(1)$; 3) The last eigenvalues decay sufficiently fast in Eq. (44) in order to avoid pathological cases.

For generic kernels, we have typically $\xi_{q,\ell} \asymp q^{-\ell}$ (for fix ℓ). For example, if h is smooth, $\xi_{q,\ell} = q^{-\ell}(h^{(k)}(0) + o_q(1))$ and it is sufficient to have $h^{(k)}(0) > 0$. See Appendix D.2 in [38] for a proof of Eq. (44) when h is sufficiently smooth.

Below, we present instead sufficient conditions on the activation σ such that the induced neural tangent kernel verifies the 'genericity' assumption. More precisely, we display sufficient conditions on the sequence $\{\sigma_q\}_{q\ge1}$ of activation functions $\sigma_q : \mathbb{R} \to \mathbb{R}$, such that the induced neural tangent kernels $\{h_q\}_{q\ge1}$ verifies Assumption 1, where h_q was derived in Section A.2 and is given by ($\boldsymbol{u},\boldsymbol{v} \in \mathscr{Q}^q$)

$$h_q(\langle \boldsymbol{u},\boldsymbol{v}\rangle/q) := h_q^{(1)}(\langle \boldsymbol{u},\boldsymbol{v}\rangle/q) + h_q^{(2)}(\langle \boldsymbol{u},\boldsymbol{v}\rangle/q)\,, \tag{53}$$

where

$$h_q^{(1)}(\langle \boldsymbol{u},\boldsymbol{v}\rangle/q) = \mathbb{E}_{\boldsymbol{w}\sim \mathrm{Unif}(\mathscr{Q}^q)}\big[\sigma_q(\langle \boldsymbol{u},\boldsymbol{w}\rangle/\sqrt{q})\sigma_q(\langle \boldsymbol{v},\boldsymbol{w}\rangle/\sqrt{q})\big]\,, \tag{54}$$

$$h_q^{(2)}(\langle \boldsymbol{u},\boldsymbol{v}\rangle/np.sqrt(q)) = \mathbb{E}_{\boldsymbol{w}\sim \mathrm{Unif}(\mathscr{Q}^q)}\big[\sigma_q'(\langle \boldsymbol{u},\boldsymbol{w}\rangle/\sqrt{q})\sigma_q'(\langle \boldsymbol{v},\boldsymbol{w}\rangle/\sqrt{q})\langle \boldsymbol{u},\boldsymbol{v}\rangle\big]/q\,. \tag{55}$$

Assumption 3 (Assumptions on $\{\sigma_q\}_{q\ge1}$ at level $\mathsf{s} \in \mathbb{N}$)**.** *For $\{\sigma_q\}_{q\ge1}$ a sequence of functions $\sigma_q : \mathbb{R} \to \mathbb{R}$, we assume the following conditions to hold. There exists $\mathsf{s}' \ge 1/\delta + 2\mathsf{s} + 3$ where $\delta > 0$ verifies $q \ge d^{\delta}$, such that*

(a) The function σ_q is differentiable and there exists $c_0 > 0$ and $c_1 < 1$ independent of q, such that $|\sigma_q(x)|, |\sigma_q'(x)| \le c_0 \exp(c_1 x^2/2)$.

(b) We have

$$\min_{k\le \mathsf{s}-1} q^{\mathsf{s}-1-k} \|\mathsf{P}_k \sigma_q(\langle \boldsymbol{e},\cdot\rangle/\sqrt{q})\|_{L^2(\mathscr{Q}^q)} = \Omega_q(1)\,, \tag{56}$$

$$\min_{k\in\{\mathsf{s},\mathsf{s}+1,\mathsf{s}'\}} \|\mathsf{P}_k \sigma_q(\langle \boldsymbol{e},\cdot\rangle/\sqrt{q})\|_{L^2(\mathscr{Q}^q)} = \Omega_q(1)\,, \tag{57}$$

where $\boldsymbol{e} \in \mathscr{Q}^q$ is arbitrary.

(c) We have for a fixed $\delta > 0$

$$\max_{k=0,\dots,\mathsf{s}'} q^{\mathsf{s}'-k+1} \|\mathsf{P}_k \sigma_q(\langle \boldsymbol{e}, \cdot\rangle/\sqrt{q})\|_{L^2(\mathscr{Q}^q)} = O_q(1)\,, \tag{58}$$

$$\max_{k=0,\dots,\mathsf{s}'} q^{\mathsf{s}'-k+1} \|\mathsf{P}_k \sigma_q'(\langle \boldsymbol{e}, \cdot\rangle/\sqrt{q})\|_{L^2(\mathscr{Q}^q)} = O_q(1)\,. \tag{59}$$

Proposition 6. *Consider a sequence $\{\sigma_q\}_{q\geq 1}$ of activation functions $\sigma_q : \mathbb{R} \to \mathbb{R}$ that satisfies Assumption 3. Let $\{h_q\}_{q\geq 1}$ be the sequence of neural tangent kernels associated to $\{\sigma_q\}_{q\geq 1}$ as defined in Eq.* (53)*. Then the sequence $\{h_q\}_{q\geq 1}$ satisfies the 'genericity' Assumption 1.*

Differentiability assumption: As mentioned in the previous section, this condition is required in our proof technique to extend Theorem 8 to non-polynomial kernel functions. While we believe that weaker conditions should be sufficient, we leave checking them to future work. Note that Assumption 2 was proved for $\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and $h_q(\langle \boldsymbol{x}, \boldsymbol{y}\rangle/q) = \mathbb{E}_{\boldsymbol{w}}\{\sigma(\langle \boldsymbol{x}, \boldsymbol{w}\rangle)\sigma(\langle \boldsymbol{y}, \boldsymbol{w}\rangle)\}$ for $\boldsymbol{w} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(1))$, given that σ satisfies some differentiability conditions, in [39].

C.3 Proof of Proposition 6

Proof of Proposition 6. **Step 1. Effective activation function.**

Let us decompose both functions σ_q and σ_q' in the Gegenbauer polynomial on the hypercube basis:

$$\sigma_q(\langle \boldsymbol{u}, \boldsymbol{v}\rangle/\sqrt{q}) = \sum_{\ell=0}^{q} \chi_{q,\ell} B(\mathscr{Q}^q; \ell) Q_\ell^{(q)}(\langle \boldsymbol{u}, \boldsymbol{v}\rangle)\,, \tag{60}$$

$$\sigma_q'(\langle \boldsymbol{u}, \boldsymbol{v}\rangle/\sqrt{q}) = \sum_{\ell=0}^{q} \kappa_{q,\ell} B(\mathscr{Q}^q; \ell) Q_\ell^{(q)}(\langle \boldsymbol{u}, \boldsymbol{v}\rangle)\,, \tag{61}$$

where we recall $B(\mathscr{Q}^q; \ell) = \binom{d}{\ell}$ and (for $\boldsymbol{e} \in \mathscr{Q}^q$ arbitrary)

$$\chi_{q,\ell}(\sigma_q) = \mathbb{E}_{\boldsymbol{u}\sim \mathrm{Unif}(\mathscr{Q}^q)}\big[\sigma_q(\langle \boldsymbol{u}, \boldsymbol{e}\rangle/\sqrt{q}) Q_\ell^{(q)}(\langle \boldsymbol{u}, \boldsymbol{e}\rangle)\big],$$
$$\kappa_{q,\ell}(\sigma_q') = \mathbb{E}_{\boldsymbol{u}\sim \mathrm{Unif}(\mathscr{Q}^q)}\big[\sigma_q'(\langle \boldsymbol{u}, \boldsymbol{e}\rangle/\sqrt{q}) Q_\ell^{(q)}(\langle \boldsymbol{u}, \boldsymbol{e}\rangle)\big].$$

From the definition of $h_q^{(1)}$ in Eq. (54) and the eigendecomposition (60), we have

$$h_q^{(1)}(\langle \boldsymbol{u}, \boldsymbol{v}\rangle/q) = \sum_{\ell=0}^{q} \chi_{q,\ell}^2 B(\mathscr{Q}^q; \ell) Q_\ell^{(q)}(\langle \boldsymbol{u}, \boldsymbol{v}\rangle).$$

Similarly, from the definition of $h_q^{(2)}$ in Eq. (55), the eigendecomposition (61) and using Lemma 1 stated below, we get

$$h_q^{(2)}(\langle \boldsymbol{u}, \boldsymbol{v}\rangle/q) = \sum_{\ell=0}^{q} \kappa_{q,\ell}^2 B(\mathscr{Q}^q; \ell) Q_\ell^{(q)}(\langle \boldsymbol{u}, \boldsymbol{v}\rangle)\langle \boldsymbol{u}, \boldsymbol{v}\rangle/q = \sum_{\ell=0}^{q} \zeta_{q,\ell}^2 B(\mathscr{Q}^q; \ell) Q_\ell^{(q)}(\langle \boldsymbol{u}, \boldsymbol{v}\rangle),$$

where

$$\zeta_{q,\ell}^2 = \frac{\ell}{q}\kappa_{q,\ell-1}^2 + \frac{q-\ell}{q}\kappa_{q,\ell+1}^2. \tag{62}$$

We can therefore define $\pi_{q,\ell} = \sqrt{\chi_{q,\ell}^2 + \zeta_{q,\ell}^2}$ and $\sigma_{\mathrm{eff},q}(\langle \cdot, \cdot\rangle/\sqrt{q}) : \mathscr{Q}^q \times \mathscr{Q}^q \to \mathbb{R}$ by

$$\sigma_{\mathrm{eff},q}(\langle \boldsymbol{u}, \boldsymbol{v}\rangle/\sqrt{q}) = \sum_{\ell=0}^{q} \pi_{q,\ell} B(\mathscr{Q}^q; \ell) Q_\ell^{(q)}(\langle \boldsymbol{u}, \boldsymbol{v}\rangle),$$

such that the NT kernel (53) can be written as the kernel of the effective activation $\sigma_{\mathrm{eff},q}$:

$$\begin{aligned} h_q(\langle \boldsymbol{u}, \boldsymbol{v}\rangle/q) &= \mathbb{E}_{\boldsymbol{\theta}\sim \mathrm{Unif}(\mathscr{Q}^d)}\Big[\sigma_{\mathrm{eff},q}(\langle \boldsymbol{u}, \boldsymbol{\theta}\rangle/\sqrt{q})\sigma_{\mathrm{eff},q}(\langle \boldsymbol{y}_{(k)}, \boldsymbol{\theta}\rangle/\sqrt{q})\Big] \\ &= \sum_{\ell=0}^{q} \pi_{q,\ell}^2 B(\mathscr{Q}^q; \ell) Q_\ell^{(q)}(\langle \boldsymbol{u}, \boldsymbol{v}\rangle)\,. \end{aligned} \tag{63}$$

We will show that h_q with Gegenbauer coefficients $\xi_{q,\ell} := \pi^2_{q,\ell}$ verifies Assumption 1.

Step 2. Decay of the eigenvalues.

Recall that the sequence $\{\sigma_q\}_{q\geq 1}$ satisfies Assumption 3 at level s. From Assumption 3.(a) (for example by adapting the proof of Lemma C.1 in [24] to the hypercube), there exists $C > 0$ such that

$$h_q(1) = \|\sigma_{\mathrm{eff},q}\|^2_{L^2(\mathscr{Q}^q)} = h_q^{(1)}(1) + h_q^{(2)}(1) = \|\sigma_q\|^2_{L^2(\mathscr{Q}^q)} + \|\sigma'_q\|^2_{L^2(\mathscr{Q}^q)} \leq C,$$

and we deduce that $\chi^2_{q,\ell}, \kappa^2_{q,\ell}, \pi^2_{q,\ell} = O_q(B(\mathscr{Q}^q;\ell)^{-1})$. Using that $B(\mathscr{Q}^q;\ell) = \binom{q}{\ell}$, we deduce that for any fixed ℓ, $\chi^2_{q,\ell}, \kappa^2_{q,\ell}, \pi^2_{q,\ell} = O_q(q^{-\ell})$. Furthermore, from Assumption 3.(c), we have for $k = 0, \ldots, \mathsf{s}' + 1$,

$$\chi^2_{q,q-k} = B(\mathscr{Q}^q; q-k)^{-1}\|\mathsf{P}_{q-k}\sigma_q\|^2_{L^2(\mathscr{Q}^q)} = O_q(q^{-\mathsf{s}'-1}),$$
$$\kappa^2_{q,q-k} = B(\mathscr{Q}^q; q-k)^{-1}\|\mathsf{P}_{q-k}\sigma'_q\|^2_{L^2(\mathscr{Q}^q)} = O_q(q^{-\mathsf{s}'-1}),$$

By Eq. (62) and the definition of $\pi^2_{q,\ell}$, we have $\pi^2_{q,q-k} = O_d(q^{-\mathsf{s}'-1})$ for any $k \leq \mathsf{s}'$, which verifies Eq. (44) in Assumption 1.

Furthermore, by Assumption 3.(b), using that $\chi^2_{q,k} = B(\mathscr{Q}^q;k)^{-1}\|\mathsf{P}_k\sigma_q\|^2_{L^2(\mathscr{Q}^q)}$ and $\xi^2_{q,k} \geq \chi^2_{q,k}$, we get

$$\min_{k\leq \mathsf{s}-1} \xi^2_{q,k} = \Omega_q(q^{-\mathsf{s}+1}),$$

and

$$\xi^2_{q,\mathsf{s}} = \Omega_q(q^{-\mathsf{s}}), \qquad \xi^2_{q,\mathsf{s}+1} = \Omega_q(q^{-\mathsf{s}-1}), \qquad \xi^2_{q,\ell'} = \Omega_q(q^{-\ell'}).$$

In particular, this implies that $\|\sigma_{\mathrm{eff},d,>\mathsf{s}}\|^2_{L^2(\mathscr{Q}^q)} \geq \|\mathsf{P}_{\mathsf{s}'}\sigma_q\|^2_{L^2(\mathscr{Q}^q)} = \Omega_q(1)$. □

Lemma 1. *Let ℓ be an integer such that $0 \leq \ell \leq q$. Consider the following Gegenbauer polynomial defined on the q-dimensional hypercube (see Section D): for $\boldsymbol{x}, \boldsymbol{y} \in \mathscr{Q}^q$,*

$$Q^{(q)}_\ell(\langle \boldsymbol{x}, \boldsymbol{y}\rangle) = \frac{1}{B(\mathscr{Q}^q;\ell)} \sum_{S\subset[q],|S|=\ell} Y_S(\boldsymbol{x})Y_S(\boldsymbol{y}),$$

where we recall the definition of the homogeneous polynomial $Y_S(\boldsymbol{x}) = \boldsymbol{x}^S = \prod_{i\in S} x_i$. We have

$$Q^{(q)}_\ell(\langle \boldsymbol{x}, \boldsymbol{y}\rangle)\langle \boldsymbol{x}, \boldsymbol{y}\rangle/q = \frac{\ell}{q} Q^{(q)}_{\ell-1}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle) + \frac{q-\ell}{q} Q^{(q)}_{\ell+1}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle),$$

with the convention $Q^{(q)}_{-1} = Q^{(q)}_{q+1} = 0$.

Proof of Lemma 1. Consider $1 \leq \ell \leq q-1$. We have

$$Q^{(q)}_\ell(\langle \boldsymbol{x}, \boldsymbol{y}\rangle)\langle \boldsymbol{x}, \boldsymbol{y}\rangle/q = \frac{1}{qB(\mathscr{Q}^q;\ell)} \sum_{S\subset[q],|S|=\ell} \sum_{i\in[q]} Y_S(\boldsymbol{x})x_i \cdot Y_S(\boldsymbol{y})y_i.$$

We have $Y_S(\boldsymbol{x})x_i = Y_{S\cup\{i\}}(\boldsymbol{x})$ if $i \notin S$, and $Y_S(\boldsymbol{x})x_i = Y_{S\setminus\{i\}}(\boldsymbol{x})$ if $i \in S$. Hence, the above sum contains sets of size $\ell-1$ and $\ell+1$. For each set $S \subset [q]$ with $|S| = \ell-1$, there $q+1-\ell$ sets $|\tilde{S}| = \ell$, such that by removing one element we can obtain S. For each set $S \subset [q]$ with $|S| = \ell+1$, there $\ell+1$ sets $|\tilde{S}| = \ell$, such that by adding one element we can obtain S.

We deduce that

$$\begin{aligned} &Q^{(q)}_\ell(\langle \boldsymbol{x}, \boldsymbol{y}\rangle)\langle \boldsymbol{x}, \boldsymbol{y}\rangle/q \\ &= \frac{q+1-\ell}{qB(\mathscr{Q}^q;\ell)} \sum_{S\subset[q],|S|=\ell-1} Y_S(\boldsymbol{x})Y_S(\boldsymbol{y}) + \frac{\ell+1}{qB(\mathscr{Q}^q;\ell)} \sum_{S\subset[q],|S|=\ell+1} Y_S(\boldsymbol{x})Y_S(\boldsymbol{y}). \end{aligned}$$

Using $B(\mathscr{Q}^q;\ell) = \binom{q}{\ell}$, we obtain

$$Q^{(q)}_\ell(\langle \boldsymbol{x}, \boldsymbol{y}\rangle)\langle \boldsymbol{x}, \boldsymbol{y}\rangle/q = \frac{\ell}{q} Q^{(q)}_{\ell-1}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle) + \frac{q-\ell}{q} Q^{(q)}_{\ell+1}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle).$$

The cases $\ell = 0$ and $\ell = q$ are straightforward. □

C.4 Proof of Theorem 7

Let $\{d(q)\}_{q\geq 1}$ be a sequence of integers with $2q \leq d(q) \leq q^{1/\delta}$ for some $\delta > 0$. We will denote $d = d(q)$ for simplicity. Consider $\boldsymbol{x} \sim \mathrm{Unif}(\mathscr{Q}^d)$, $dq^{\mathsf{s}-1+\delta} \leq n \leq dq^{\mathsf{s}-\delta}$ for some $\delta > 0$ and a sequence of inner-product kernels $\{h_q\}_{q\geq 1}$ that satisfies Assumption 1 at level s. We consider the vanilla one-layer convolutional kernel

$$H^{\mathsf{CK},d}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{d}\sum_{k=1}^{d} h_q(\langle \boldsymbol{x}_{(k)}, \boldsymbol{y}_{(k)}\rangle/q).$$

Theorem 7 is a consequence of Theorem 4 in [38] where we take $\mathcal{X}_d = \mathscr{Q}^d$, $\nu_d = \mathrm{Unif}(\mathcal{X}_d)$ and $\mathcal{D}_d = L^2(\mathscr{Q}^d, \mathrm{Loc}_q) \subset L^2(\mathscr{Q}^d)$. The proof amounts to checking that $\{H^{\mathsf{CK},d}\}_{q\geq 1}$ verifies the kernel concentration properties and eigenvalue condition (see Section 3.2 in [38]). We borrow some of the notations introduced in [38] and we refer the reader to their Section 2.1.

Proof of Theorem 7. **Step 1. Diagonalization of the kernel and choosing $\mathsf{m} = \mathsf{m}(q)$.**

From Proposition 1, we have the following diagonalization of $H^{\mathsf{CK},d}$:

$$H_d(\boldsymbol{x},\boldsymbol{y}) := H^{\mathsf{CK},d}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{d}\sum_{\ell=0}^{q}\sum_{S\in\mathcal{E}_\ell} \xi_{q,\ell} r(S)\cdot Y_S(\boldsymbol{x})Y_S(\boldsymbol{y}),$$

where $r(\emptyset) = d$ and $r(S) = q+1-\gamma(S)$ for $S \subset [q]\setminus\{\emptyset\}$, and we recall $\mathcal{E}_\ell = \{S \subseteq [d] : |S| = \ell, \gamma(S) \leq q\}$. Using that $B(\mathscr{Q}^q;\ell) = \Theta_q(q^\ell)$, $\xi_{q,\ell}B(\mathscr{Q}^q;\ell) \leq h_q(1)$ and Assumption 1, we have

$$\begin{aligned} \min_{\ell\leq \mathsf{s}-1} \xi_{q,\ell} &= \Omega_q(q^{-\mathsf{s}+1}), & \xi_{q,\mathsf{s}} &= \Theta_q(q^{-\mathsf{s}}), \\ \xi_{q,\mathsf{s}+1} &= \Theta_q(q^{-\mathsf{s}-1}), & \sup_{\ell\geq \mathsf{s}+2} \xi_{q,\ell} &= O_q(q^{-\mathsf{s}-2}). \end{aligned} \tag{64}$$

Further define $\mathcal{E}_{\ell,h} = \{S \in \mathcal{E}_\ell : \gamma(S) = h\}$ for $h = \ell,\dots,q$. It is easy to check that $|\mathcal{E}_{\ell,h}| = d\binom{h-2}{\ell-2}$ and

$$|\mathcal{E}_\ell| = \sum_{h=\ell}^{q} |\mathcal{E}_{\ell,h}| = d\sum_{h=\ell}^{q}\binom{h-2}{\ell-2} = d\binom{q-1}{\ell-1},$$

and therefore $|\mathcal{E}_\ell| = \Theta_q(d\cdot q^{\ell-1})$.

Denote $\{\lambda_{q,j}\}_{j\geq 1}$ the eigenvalues $\{\xi_{q,\ell} r(S)/d\}_{\ell=0,\dots,q; S\in\mathcal{E}_\ell}$ in nonincreasing order, and $\{\psi_{q,j}\}_{j\geq 1}$ the reordered eigenfunctions. Set m to be the number of eigenvalues such that $\lambda_{q,j} > q\xi_{q,\mathsf{s}+1}/d$ (recall $q\xi_{q,\mathsf{s}+1} = \Theta_d(q^{-\mathsf{s}})$). Denote $\alpha = q\xi_{q,\mathsf{s}+1}/\xi_{q,\mathsf{s}}$. From the bounds (64) on $\xi_{q,\mathsf{s}+1}$ and $\xi_{q,\mathsf{s}}$, we have $\alpha = \Theta_q(1)$. Denote $\tilde{\alpha} = q+1-\alpha$ and $\mathcal{E}_{\mathsf{s},\geq\tilde{\alpha}} = \{S \in \mathcal{E}_\mathsf{s} : \gamma(S) \geq \tilde{\alpha}\}$ and $\mathcal{E}_{\mathsf{s},<\tilde{\alpha}} = \mathcal{E}_\mathsf{s} \setminus \mathcal{E}_{\mathsf{s},\geq\tilde{\alpha}}$. Using Eq. (64) and that $1 \leq r(S) \leq q$, we have $\{\lambda_{d,j}\}_{j\in[\mathsf{m}]}$ that contains exactly the eigenvalues associated to homogeneous polynomials of degree less or equal to $\mathsf{s}-1$ and of degree s with $S \in \mathcal{E}_{\mathsf{s},<\tilde{\alpha}}$ (which corresponds to the sets S such that $r(S) > \alpha$, i.e., $\xi_{q,\mathsf{s}} r(S) > q\xi_{q,\mathsf{s}+1}$). In particular, if $\alpha < 1$, then $\{\lambda_{d,j}\}_{j\in[\mathsf{m}]}$ contains exactly the eigenvalues associated to all homogeneous polynomials of degree less or equal to s.

Note that we have

$$\mathsf{m} \leq \sum_{\ell=0}^{\mathsf{s}} |\mathcal{E}_\ell| = O_q(dq^{\mathsf{s}-1}) = O_q(q^{-\delta}n). \tag{65}$$

Step 2. Diagonal elements of the truncated kernel.

Define the truncated kernel $H_{d,>\mathsf{m}}$ to be

$$\begin{aligned} H_{d,>\mathsf{m}}(\boldsymbol{x},\boldsymbol{y}) &= \sum_{j\geq \mathsf{m}+1} \lambda_{q,j}\psi_{q,j}(\boldsymbol{x})\psi_{d,j}(\boldsymbol{y}) \\ &= \frac{\xi_{q,\mathsf{s}}}{d}\sum_{S\in\mathcal{E}_{\mathsf{s},\geq\tilde{\alpha}}} r(S)\cdot Y_S(\boldsymbol{x})Y_S(\boldsymbol{y}) + \frac{1}{d}\sum_{\ell=\mathsf{s}+1}^{q}\xi_{q,\ell}\sum_{S\in\mathcal{E}_\ell} r(S)\cdot Y_S(\boldsymbol{x})Y_S(\boldsymbol{y}). \end{aligned}$$

The diagonal elements of the truncated kernel are given by: for any $\boldsymbol{x} \in \mathscr{Q}^d$,

$$H_{d,>\mathsf{m}}(\boldsymbol{x},\boldsymbol{x}) = \frac{\xi_{q,\mathsf{s}}}{d} \sum_{S \in \mathcal{E}_{\mathsf{s},\geq \tilde{\alpha}}} r(S) + \frac{1}{d} \sum_{\ell=\mathsf{s}+1}^{q} \xi_{q,\ell} \sum_{S \in \mathcal{E}_\ell} r(S) = \mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}}). \tag{66}$$

Notice that

$$\sum_{S \in \mathcal{E}_\ell} r(S) = \sum_{h=\ell}^{q} (q+1-h)|\mathcal{E}_{\ell,h}| = d \sum_{h=\ell}^{q} (q+1-h) \binom{h-2}{\ell-2} = d \binom{q}{\ell} = dB(\mathscr{Q}^q;\ell),$$

$$\sum_{S \in \mathcal{E}_{\mathsf{s},\geq \tilde{\alpha}}} r(S) \leq \alpha \sum_{h=q+1-\alpha}^{q} |\mathcal{E}_{\mathsf{s},h}| \leq d\alpha^2 \binom{q-2}{\mathsf{s}-2} = O_d(dq^{\mathsf{s}-2}).$$

Hence using that $\xi_{q,\mathsf{s}} = O_d(q^{-\mathsf{s}})$, we have

$$\mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}}) = \frac{\xi_{q,\mathsf{s}}}{d} \sum_{S \in \mathcal{E}_{\mathsf{s},\geq \tilde{\alpha}}} r(S) + \sum_{\ell=\mathsf{s}+1}^{q} \xi_{q,\ell} B(\mathscr{Q}^q;\ell) = h_{q,>\mathsf{s}}(1) + o_{q,\mathbb{P}}(1),$$

where $h_{q,>\mathsf{s}}$ is the inner-product kernel with the $(\mathsf{s}+1)$-first Gegenbauer coefficients set to zero, i.e., $h_{q,>\mathsf{s}}(\langle \boldsymbol{u},\boldsymbol{v}\rangle/q) = \sum_{\ell=s+1}^{q} \xi_{q,\ell} B(\mathscr{Q}^q;\ell) Q_\ell^{(q)}(\langle \boldsymbol{u},\boldsymbol{v}\rangle)$, for any $\boldsymbol{u},\boldsymbol{v} \in \mathscr{Q}^q$. From Assumption 1 at level s, we have $\Omega_q(1) = \xi_{q,\ell'} B(\mathscr{Q}^q;\ell') \leq h_{q,>\mathsf{s}}(1) \leq h_q(1) = O_q(1)$. Hence, $\mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}}) = \Theta_d(1)$.

Similarly,

$$\mathbb{E}_{\boldsymbol{x}'}[H_{d,>\mathsf{m}}(\boldsymbol{x},\boldsymbol{x}')^2] = \frac{\xi_{q,\mathsf{s}}^2}{d} \sum_{S \in \mathcal{E}_{\mathsf{s},\geq \tilde{\alpha}}} r(S)^2 + \frac{1}{d} \sum_{\ell=\mathsf{s}+1}^{q} \xi_{q,\ell}^2 \sum_{S \in \mathcal{E}_\ell} r(S)^2 = \mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}}^2). \tag{67}$$

Step 3. Choosing the sequence $u = u(d)$.

Let s' be chosen as in Assumption 1, i.e., such that $\xi_{q,\mathsf{s}'} B(\mathscr{Q}^q;\mathsf{s}') = \Omega_q(1)$. We have

$$\xi_{q,\mathsf{s}'} = \Theta_q(q^{-\mathsf{s}'}), \qquad \sup_{\ell \geq \mathsf{s}'+1} \xi_{q,\ell} = O_q(q^{-\mathsf{s}'-1}). \tag{68}$$

Set $u = u(d)$ to be the number of eigenvalues such that $\lambda_{q,j} > q\xi_{q,\mathsf{s}'}/d = \Theta_q(q^{-\mathsf{s}'+1}/d)$. From Eqs. (64) and (68), and recalling that $1 \leq r(S) \leq q$, we deduce that $\{\lambda_{d,j}\}_{j\in[u]}$ must contain all the eigenvalues associated to homogeneous polynomials of degree less or equal to ℓ and does not contain any of the eigenvalues associated to homogeneous polynomials of degree larger or equal to s'.

We have

$$\mathrm{Tr}(\mathbb{H}_{d,>u}) = \sum_{j>u} \lambda_{q,j} \leq \mathrm{Tr}(\mathbb{H}_{d,>m}) = O_q(1),$$

$$\mathrm{Tr}(\mathbb{H}_{d,>u}) \geq \frac{\xi_{q,\mathsf{s}'}}{d} \sum_{S \in \mathcal{E}_{\mathsf{s}'}} r(S) = \xi_{q,\mathsf{s}'} B(\mathscr{Q}^q;\mathsf{s}') = \Omega_q(1).$$

Similarly, we have

$$\mathrm{Tr}(\mathbb{H}_{d,>u}^2) = \sum_{j>u} \lambda_{q,j}^2 \leq \mathrm{Tr}(\mathbb{H}_{d,>u}) \cdot \sup_{j>m} \lambda_{d,j} = qd^{-1}\xi_{q,\mathsf{s}'} \mathrm{Tr}(\mathbb{H}_{d,>m}) = O_q(d^{-1}q^{-\mathsf{s}'+1}),$$

$$\mathrm{Tr}(\mathbb{H}_{d,>u}^2) \geq \frac{\xi_{q,\mathsf{s}'}^2}{d^2} \sum_{S \in \mathcal{E}_{\mathsf{s}'}} r(S)^2 \geq d^{-1}\xi_{q,\mathsf{s}'}^2 B(\mathscr{Q}^q;\mathsf{s}') = \Omega_q(d^{-1}q^{-\mathsf{s}'}).$$

Finally,

$$\mathrm{Tr}(\mathbb{H}_{d,>u}^4) = \sum_{j>u} \lambda_{d,j}^4 \leq d^{-3}q^3\xi_{q,\mathsf{s}'}^3 \mathrm{Tr}(\mathbb{H}_{d,>m}) = O_q(d^{-3}q^{-3\ell'+3}).$$

Step 4. Checking the kernel concentration property at level $\{(n(q),\mathsf{m}(q))\}_{q\geq 1}$.

Let us check the kernel concentration property at level (n,m) with the sequence of integers $\{u(q)\}_{q\geq 1}$ defined in the previous step (Assumption 4 in [38]):

(a) *(Hypercontractivity of finite eigenspaces)* The subspace spanned by the top eigenvectors $\{\psi_{q,j}\}_{j\in[u]}$ is contained in the subspace of polynomials of degree less or equal to $\mathsf{s}'-1$ on the hypercube. The hypercontractivity of this subspace is a consequence of a classical result due to Beckner, Bonami and Gross (see Lemma 4 in Section D).

(b) *(Properly decaying eigenvalues.)* From step 3 and recalling that $\mathsf{s}' \geq 1/\delta + 2\mathsf{s} + 3$ where $\delta > 0$ verifies $q \geq d^{\delta}$, we have

$$\frac{\mathrm{Tr}(\mathbb{H}_{d,>u})^2}{\mathrm{Tr}(\mathbb{H}_{d,>u}^2)} = \Omega_q(1)\cdot dq^{\mathsf{s}'-1} = \Omega_q(1)\cdot d^2 q^{2\mathsf{s}+1} \geq n^{2+\delta'},$$

for $\delta' > 0$ sufficiently small. Similarly,

$$\frac{\mathrm{Tr}(\mathbb{H}_{d,>u}^2)^2}{\mathrm{Tr}(\mathbb{H}_{d,>u}^4)} = \Omega_q(1)\cdot dq^{\mathsf{s}'-3} = \Omega_q(1)\cdot d^2 q^{2\mathsf{s}} \geq n^{2+\delta'},$$

for $\delta' > 0$ chosen sufficiently small.

(c) *(Concentration of the diagonal elements of the kernel)* From Eqs. (66) and (67), the diagonal elements of the kernel are constant and the assumption is automatically verified.

Step 5. Checking the eigenvalue condition at level $\{(n(q), \mathsf{m}(q))\}_{q\geq 1}$.

Let us now check the eigenvalue condition at level $\{(n(q), \mathsf{m}(q))\}_{q\geq 1}$ which corresponds to Assumption 5 in [38]):

(a) First notice that

$$\begin{aligned}\sum_{S\in\mathcal{E}_{\mathsf{s}+1}} r(S)^2 &= d\sum_{h=\mathsf{s}+1}^{q}(q+1-h)^2\binom{h-1}{\mathsf{s}-1} \geq d\sum_{h=\mathsf{s}+1}^{\lfloor q/2\rfloor}(q+1-h)^2\binom{h-1}{\mathsf{s}-1} \\ &\geq \frac{dq^2}{4}\sum_{h=\mathsf{s}+1}^{\lfloor q/2\rfloor}\binom{h-1}{\mathsf{s}-1} = \frac{dq^2}{4}\binom{\lfloor q/2\rfloor}{\mathsf{s}} = \Omega_q(1)\cdot dq^{2+\mathsf{s}}.\end{aligned} \tag{69}$$

Hence

$$\frac{\mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}}^2)}{\lambda_{d,\mathsf{m}+1}^2} \geq \frac{\sum_{S\in\mathcal{E}_{\mathsf{s}+1}}\xi_{d,\mathsf{s}+1}^2 r(S)^2}{q^2\xi_{d,\mathsf{s}+1}^2} = \Omega_q(1)\cdot dq^{\mathsf{s}} \geq n^{1+\delta},$$

for $\delta > 0$ sufficiently small. Similarly,

$$\frac{\mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}})}{\lambda_{d,\mathsf{m}+1}} = \Omega_q(1)\cdot\frac{d}{q\xi_{d,\mathsf{s}+1}} = \Omega_d(1)\cdot dq^{\mathsf{s}} \geq n^{1+\delta}.$$

(b) This is a direct consequence of Eq. (65).

We can therefore apply Theorem 4 in [38], which concludes the proof. □

C.5 Proof of Theorem 8

Consider $q^{\mathsf{s}-1+\delta} \leq n \leq q^{\mathsf{s}-\delta}$ for some $\delta > 0$ and a sequence of inner-product kernels $\{h_q\}_{q\geq 1}$ that satisfies Assumptions 1 and 2 at level s. We consider the one-layer convolutional kernel with global average pooling

$$H_{\mathsf{GP}}^{\mathsf{CK},d}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{d}\sum_{k,k'=1}^{d} h_q\big(\langle \boldsymbol{x}_{(k)}, \boldsymbol{y}_{(k')}\rangle/q\big).$$

Again, the proof of Theorem 8 will amount to checking that the conditions of Theorem 4 in [38] hold.

For the sake of simplicity, we will further assume that $\xi_{q,\mathsf{s}} > q\xi_{q,\mathsf{s}+1}$, which simplifies some of the computation. This condition can be removed as in Theorem 7, by considering the set $\mathcal{C}_{\mathsf{s},<\tilde{\alpha}} = \{S \in \mathcal{C}_{\mathsf{s}} : \gamma(S) < \tilde{\alpha}\}$ and showing that the extra terms corresponding to these eigenfunctions are negligible.

Proof of Theorem 8. **Step 1. Diagonalization of the kernel and choosing $\mathsf{m} = \mathsf{m}(q)$.**

From Proposition 2 with $\omega = d$, we have the following diagonalization of $H^d_{d,q}$:

$$H_d(\boldsymbol{x}, \boldsymbol{y}) := H^{\mathsf{CK},d}_{\mathsf{GP}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\ell=0}^{q} \sum_{S \in \mathcal{C}_\ell} \xi_{q,\ell} r(S) \cdot \psi_S(\boldsymbol{x}) \psi_S(\boldsymbol{y}),$$

where we recall $\psi_S(\boldsymbol{x}) = \frac{1}{\sqrt{d}} \sum_{k \in [d]} Y_{k+S}(\boldsymbol{x})$ and that $\mathcal{C}_\ell$ is the quotient space of $\mathcal{E}_\ell$ with the translation equivalence relation. It is easy to check that $|\mathcal{C}_\ell| = \binom{q-1}{\ell-1}$.

From Assumption 1, we get the same bounds on the Gegenbauer coefficients $\xi_{q,\ell}$ as Eq. (64) in the proof of Theorem 7. Denote $\{\lambda_{q,j}\}_{j \geq 1}$ the eigenvalues $\{\xi_{q,\ell} r(S)\}_{\ell=0,\dots,q; S \in \mathcal{E}_\ell}$ in nonincreasing order, and $\{\psi_{q,j}\}_{j \geq 1}$ the reordered eigenfunctions. Set m to be the number of eigenvalues such that $\lambda_{q,j} > q\xi_{q,\mathsf{s}+1}$ (recall $q\xi_{q,\mathsf{s}+1} = \Theta_d(q^{-\mathsf{s}})$). From the bounds (64) and our simplifying assumption that $\xi_{q,\mathsf{s}} > q\xi_{q,\mathsf{s}+1}$, we have $\{\lambda_{d,j}\}_{j \in [\mathsf{m}]}$ that contains exactly the eigenvalues associated to homogeneous polynomials of degree less or equal to s.

Note that we have

$$\mathsf{m} = \sum_{\ell=0}^{\mathsf{s}} |\mathcal{C}_\ell| = O_q(q^{\mathsf{s}-1}) = O_q(q^{-\delta} n). \tag{70}$$

Step 2. Diagonal elements of the truncated kernel.

Define the truncated kernel $H_{d,>\mathsf{m}}$ to be

$$H_{d,>\mathsf{m}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j \geq \mathsf{m}+1} \lambda_{d,j} \psi_{d,j}(\boldsymbol{x}) \psi_{d,j}(\boldsymbol{y}) = \sum_{\ell=\mathsf{s}+1}^{q} \sum_{S \in \mathcal{C}_\ell} \xi_{q,\ell} r(S) \cdot \psi_S(\boldsymbol{x}) \psi_S(\boldsymbol{y}).$$

The diagonal elements of the truncated kernel are given by: for any $\boldsymbol{x} \in \mathscr{Q}^d$,

$$H_{d,>\mathsf{m}}(\boldsymbol{x}, \boldsymbol{x}) = \sum_{\ell=\mathsf{s}+1}^{q} \xi_{q,\ell} B(\mathscr{Q}^q; \ell) \Upsilon^{(q)}_\ell(\boldsymbol{x}),$$

where

$$\Upsilon^{(q)}_\ell(\boldsymbol{x}) = \frac{1}{B(\mathscr{Q}^q; \ell)} \sum_{S \in \mathcal{C}_\ell} r(S) \psi_S(\boldsymbol{x})^2.$$

Notice that we have now

$$\sum_{S \in \mathcal{C}_\ell} r(S) = \sum_{h=\ell}^{q} (q+1-h) \binom{h-2}{\ell-2} = \binom{q}{\ell} = B(\mathscr{Q}^q; \ell).$$

Therefore $\mathbb{E}_{\boldsymbol{x}}[\Upsilon^{(q)}_\ell(\boldsymbol{x})] = 1$ and

$$\mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}}) = \mathbb{E}_{\boldsymbol{x}}[H_{d,>\mathsf{m}}(\boldsymbol{x}, \boldsymbol{x})] = \sum_{\ell=\mathsf{s}+1}^{q} \xi_{q,\ell} B(\mathscr{Q}^q; \ell) = h_{q,>\mathsf{s}}(1).$$

From Proposition 7 with $\ell = \mathsf{s}$, we have

$$\begin{aligned} \sup_{i \in [n]} \Big| H_{d,>\mathsf{m}}(\boldsymbol{x}_i, \boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}}[H_{d,>\mathsf{m}}(\boldsymbol{x}, \boldsymbol{x})] \Big| &= \mathrm{Tr}(\mathbb{H}_{d,>\mathsf{m}}) \cdot o_{d,\mathbb{P}}(1), \\ \sup_{i \in [n]} \Big| \mathbb{E}_{\boldsymbol{x}'}[H_{d,>\mathsf{m}}(\boldsymbol{x}_i, \boldsymbol{x}')^2] - \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}[H_{d,>\mathsf{m}}(\boldsymbol{x}, \boldsymbol{x}')^2] \Big| &= \mathrm{Tr}(\mathbb{H}^2_{d,>\mathsf{m}}) \cdot o_{d,\mathbb{P}}(1). \end{aligned} \tag{71}$$

Step 3. Choosing the sequence $u = u(d)$.

Let s' be chosen as in Assumption 1. Similarly to step 3 in the proof of Theorem 7, take $u = u(d)$ to be the number of eigenvalues such that $\lambda_{q,j} > q\xi_{q,\ell'}$. We get

$$\begin{aligned} \mathrm{Tr}(\mathbb{H}_{d,>u}) &= \Theta_q(1), \\ \mathrm{Tr}(\mathbb{H}^2_{d,>u}) &= O_q(q^{-\mathsf{s}'+1}), \\ \mathrm{Tr}(\mathbb{H}^2_{d,>u}) &= \Omega_q(q^{-\ell'}), \\ \mathrm{Tr}(\mathbb{H}^4_{d,>u}) &= O_q(q^{-3\mathsf{s}'+3}). \end{aligned}$$

Step 4. Checking the kernel concentration property at level $\{(n(q), \mathsf{m}(q))\}_{q\geq 1}$.

The kernel concentration property at level (n, m) hold with the sequence $\{u(q)\}_{q\geq 1}$ as defined in step 3. The hypercontractivity of finite eigenspaces and the properly decaying eigenvalues are obtained as in step 4 of the proof of Theorem 7, while the concentration of the diagonal elements of the kernel is given by Eq. (71).

Step 5. Checking the eigenvalue condition at level $\{(n(q), \mathsf{m}(q))\}_{q\geq 1}$.

This is obtained similarly as in step 5 of the proof of Theorem 7.

□

C.6 Auxiliary results

Proposition 7. *Let $\mathsf{s} \geq 1$ be a fixed integer. Assume that the sequence of inner-product kernels $\{h_q\}_{q\geq 1}$ satisfies Assumptions 1 and 2 at level s. Define $H_d^{>\mathsf{s}} : \mathscr{Q}^d \times \mathscr{Q}^d \to \mathbb{R}$ as the convolutional kernel with global average pooling*

$$H_d^{>\mathsf{s}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{d} \sum_{k,k'\in[d]} h_{q,>\mathsf{s}}(\langle \boldsymbol{x}_{(k)}, \boldsymbol{y}_{(k')}\rangle/q),$$

where $h_{q,>\mathsf{s}}$ is the inner-product kernel where the $\mathsf{s}+1$ first Gegenbauer coefficients are set to 0.

Then for $n = O_q(q^p)$ for some fixed p, letting $(\boldsymbol{x}_i)_{i\in[n]} \sim \mathrm{Unif}(\mathscr{Q}^d)$, we have

$$\sup_{i\in[n]} \left| H_d^{>\mathsf{s}}(\boldsymbol{x}_i, \boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}}[H_d^{>\mathsf{s}}(\boldsymbol{x}, \boldsymbol{x})] \right| = \mathbb{E}_{\boldsymbol{x}}[H_d^{>\mathsf{s}}(\boldsymbol{x}, \boldsymbol{x})] \cdot o_{d,\mathbb{P}}(1), \tag{72}$$

$$\sup_{i\in[n]} \left| \mathbb{E}_{\boldsymbol{x}'}[H_d^{>\mathsf{s}}(\boldsymbol{x}_i, \boldsymbol{x}')^2] - \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}[H_d^{>\mathsf{s}}(\boldsymbol{x}, \boldsymbol{x}')^2] \right| = \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}[H_d^{>\mathsf{s}}(\boldsymbol{x}, \boldsymbol{x}')^2] \cdot o_{d,\mathbb{P}}(1). \tag{73}$$

Proof of Proposition 7. **Step 1. Bounding $\sup_{i\in[n]} \left| H_d^{>\mathsf{s}}(\boldsymbol{x}_i, \boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}}[H_d^{>\mathsf{s}}(\boldsymbol{x}, \boldsymbol{x})] \right|$.**

Recall that we defined

$$\Upsilon_\ell^{(q)}(\boldsymbol{x}) = \frac{1}{B(\mathscr{Q}^q; \ell)} \sum_{S\in\mathcal{C}_\ell} r(S)\psi_S(\boldsymbol{x})^2.$$

Following the same proof as Proposition 8 in [39], notice that for the integer v in Assumption 2, by Lemma 2 stated below, we have

$$\begin{aligned}
&\sup_{i\in[n]} \left| H_d^{>\mathsf{s}}(\boldsymbol{x}_i, \boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}}[H_d^{>\mathsf{s}}(\boldsymbol{x}, \boldsymbol{x})] \right| \\
\leq &\sup_{i\in[n]} \left| H_d^{>v}(\boldsymbol{x}_i, \boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}}[H_d^{>v}(\boldsymbol{x}, \boldsymbol{x})] \right| + \sum_{\ell=\mathsf{s}+1}^{v} \xi_{q,\ell} B(\mathscr{Q}^q; \ell) \cdot \max_{i\in[n]} \left| \Upsilon_\ell^{(d)}(\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}}[\Upsilon_\ell^{(d)}(\boldsymbol{x})] \right| \\
= &\sup_{i\in[n]} \left| H_d^{>v}(\boldsymbol{x}_i, \boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}}[H_d^{>v}(\boldsymbol{x}, \boldsymbol{x})] \right| + \left(\sum_{\ell=\mathsf{s}+1}^{v} \xi_{q,\ell} B(\mathscr{Q}^q; \ell) \right) \cdot o_{d,\mathbb{P}}(1).
\end{aligned}$$

By Assumption 2, there exists $C > 0$ such that for any $\gamma \in [-1, 1]$,

$$\left| h_{q,>v}(\gamma) - \sum_{r=0}^{v} \frac{1}{r!} h_{q,>v}^{(r)}(0)\gamma^r \right| \leq C \cdot |\gamma|^{v+1}, \tag{74}$$

and $|h_{q,>v}^{(r)}(0)| \leq Cq^{-(v+1-r)/2}$ for $r \leq v$. Moreover, by Hanson-Wright inequality as in Lemma 3, using $n = O_q(q^p)$ (at most polynomial in q) and a union bound, we have for any $\eta > 0$,

$$\sup_{1\leq r\leq v+1} \sup_{k\neq l} \sup_{i\in[n]} \left| \langle (\boldsymbol{x}_i)_{(k)}, (\boldsymbol{x}_i)_{(l)} \rangle^r \right| \cdot q^{-k/2-\eta} = o_{q,\mathbb{P}}(1),$$

$$\sup_{1\leq r\leq v+1} \sup_{k\neq l} \mathbb{E}\left[\left| \langle \boldsymbol{x}_{(k)}, \boldsymbol{x}_{(l)} \rangle^r \right| \right] \cdot q^{-k/2-\eta} = o_{q,\mathbb{P}}(1).$$

Therefore, injecting these bounds in Eq. (74), we get

$$\sup_{k\neq l}\sup_{i\in[n]}\left|h_{q,>v}(\langle(\boldsymbol{x}_i)_{(k)},(\boldsymbol{x}_i)_{(l)}\rangle/q)\right| = O_{q,\mathbb{P}}(q^{-(v+1)/2+\eta}),$$
$$\sup_{k\neq l}\mathbb{E}\left[\left|h_{q,>v}(\langle\boldsymbol{x}_{(k)},\boldsymbol{x}_{(l)}\rangle/q)\right|\right] = O_{q,\mathbb{P}}(q^{-(v+1)/2+\eta}).$$

Hence, we deduce that

$$\begin{aligned}
&\sup_{i\in[n]}\left|H_d^{>v}(\boldsymbol{x}_i,\boldsymbol{x}_i)-\mathbb{E}_{\boldsymbol{x}}[H_d^{>v}(\boldsymbol{x},\boldsymbol{x})]\right| \\
&\leq \frac{1}{d}\sum_{k\neq l\in[d]}\sup_{i\in[n]}\left|h_{q,>v}(\langle(\boldsymbol{x}_i)_{(k)},(\boldsymbol{x}_i)_{(l)}\rangle/q)-\mathbb{E}_{\boldsymbol{x}}[h_{q,>v}(\langle\boldsymbol{x}_{(k)},\boldsymbol{x}_{(l)}\rangle/q)]\right| \\
&\leq d\sup_{k\neq l}\left\{\sup_{i\in[n]}\left|h_{q,>v}(\langle(\boldsymbol{x}_i)_{(k)},(\boldsymbol{x}_i)_{(l)}\rangle/q)\right|+\mathbb{E}\left[\left|h_{q,>v}(\langle\boldsymbol{x}_{(k)},\boldsymbol{x}_{(l)}\rangle/q)\right|\right]\right\} \\
&= O_{q,\mathbb{P}}(dq^{-(v+1)/2+\eta}) = o_{d,\mathbb{P}}(1).
\end{aligned}$$

Furthermore, recall that by Assumption 1, we have $\mathbb{E}[\mathbb{H}_d^{>\ell}(\boldsymbol{x},\boldsymbol{x})]\geq\xi_{q,\mathsf{s}'}B(\mathscr{Q}^q;\mathsf{s}')=\Omega_q(1)$. We get

$$\sup_{i\in[n]}\left|H_d^{>v}(\boldsymbol{x}_i,\boldsymbol{x}_i)-\mathbb{E}_{\boldsymbol{x}}[H_d^{>v}(\boldsymbol{x},\boldsymbol{x})]\right| = \mathbb{E}[\mathbb{H}_d^{>\ell}(\boldsymbol{x},\boldsymbol{x})]\cdot o_{q,\mathbb{P}}(1),$$

which concludes the proof of the first bound.

Step 2. Bounding $\sup_{i\in[n]}\left|\mathbb{E}_{\boldsymbol{x}'}[H_d^{>\mathsf{s}}(\boldsymbol{x}_i,\boldsymbol{x}')^2]-\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}[H_d^{>\mathsf{s}}(\boldsymbol{x},\boldsymbol{x}')^2]\right|$**.**

Notice that we can write,

$$\mathbb{E}_{\boldsymbol{x}'}[H_d^{>\mathsf{s}}(\boldsymbol{x},\boldsymbol{x}')^2] = \sum_{\ell=\mathsf{s}+1}^{q}\xi_{q,\ell}^2R_\ell\cdot\Xi_\ell^{(d)}(\boldsymbol{x}),$$

where we denoted $R_\ell=\sum_{S\in\mathcal{C}_\ell}r(S)^2$ and

$$\Xi_\ell^{(d)}(\boldsymbol{x}) = \frac{1}{R_\ell}\sum_{S\in\mathcal{C}_\ell}r(S)^2\psi_S(\boldsymbol{x})^2.$$

Then, by Lemma 2, we get for any $u\geq\mathsf{s}$,

$$\begin{aligned}
&\sup_{i\in[n]}\left|\mathbb{E}_{\boldsymbol{x}'}[H_d^{>\mathsf{s}}(\boldsymbol{x}_i,\boldsymbol{x}')^2]-\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}[H_d^{>\mathsf{s}}(\boldsymbol{x},\boldsymbol{x}')^2]\right| \\
&\leq \sup_{i\in[n]}\left|\mathbb{E}_{\boldsymbol{x}'}[H_d^{>u}(\boldsymbol{x}_i,\boldsymbol{x}')^2]-\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}[H_d^{>u}(\boldsymbol{x},\boldsymbol{x}')^2]\right|+\sum_{\ell=\mathsf{s}+1}^{u}\xi_{q,\ell}^2R_\ell\cdot\max_{i\in[n]}\left|\Xi_\ell^{(d)}(\boldsymbol{x}_i)-\mathbb{E}_{\boldsymbol{x}}[\Xi_\ell^{(d)}(\boldsymbol{x})]\right| \\
&= \sup_{i\in[n]}\left|\mathbb{E}_{\boldsymbol{x}'}[H_d^{>u}(\boldsymbol{x}_i,\boldsymbol{x}')^2]-\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}[H_d^{>u}(\boldsymbol{x},\boldsymbol{x}')^2]\right|+\left(\sum_{\ell=\mathsf{s}+1}^{u}\xi_{q,\ell}^2R_\ell\right)\cdot o_{d,\mathbb{P}}(1).
\end{aligned}$$

We conclude following the same argument as in the proof of Proposition 9 in [39]. □

Lemma 2. *Let $\ell\geq 2$ be an integer. Define $\Upsilon_\ell^{(d)}:\mathscr{Q}^d\to\mathbb{R}$ and $\Xi_\ell^{(d)}:\mathscr{Q}^d\to\mathbb{R}$ to be*

$$\Upsilon_\ell^{(d)}(\boldsymbol{x}) = \frac{1}{B(\mathscr{Q}^q;\ell)}\sum_{S\in\mathcal{C}_\ell}r(S)\psi_S(\boldsymbol{x})^2, \tag{75}$$

$$\Xi_\ell^{(d)}(\boldsymbol{x}) = \frac{1}{R_\ell}\sum_{S\in\mathcal{C}_\ell}r(S)^2\psi_S(\boldsymbol{x})^2, \tag{76}$$

where $R_\ell=\sum_{S\in\mathcal{C}_\ell}r(S)^2$.

Let $n \le q^p$ for some fixed p. Then, for $(\boldsymbol{x}_i)_{i\in[n]} \overset{i.i.d.}{\sim} \mathrm{Unif}(\mathscr{Q}^d)$, we have

$$\max_{i\in[n]} \left|\Upsilon_\ell^{(d)}(\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}}[\Upsilon_\ell^{(d)}(\boldsymbol{x})]\right| = o_{d,\mathbb{P}}(1), \tag{77}$$

$$\max_{i\in[n]} \left|\Xi_\ell^{(d)}(\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}}[\Xi_\ell^{(d)}(\boldsymbol{x})]\right| = o_{d,\mathbb{P}}(1), \tag{78}$$

where $\mathbb{E}_{\boldsymbol{\theta}}[\Upsilon_\ell^{(d)}(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{x}}[\Xi_\ell^{(d)}(\boldsymbol{x})] = 1$.

Proof of Lemma 2. **Step 1. Bounding $\max_{i\in[n]} \left|\Upsilon_\ell^{(d)}(\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}}[\Upsilon_\ell^{(d)}(\boldsymbol{x})]\right|$.**

Define $F_\ell : \mathscr{Q}^d \to \mathbb{R}$ to be

$$F_\ell(\boldsymbol{x}) = \Upsilon_\ell^{(d)}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}}[\Upsilon_\ell^{(d)}(\boldsymbol{x})] = \frac{1}{dB(\mathscr{Q}^q;\ell)} \sum_{S\in\mathcal{C}_\ell} r(S) \sum_{i\neq j\in[d]} Y_{i+S}(\boldsymbol{x})Y_{j+S}(\boldsymbol{x}). \tag{79}$$

Notice that $F_\ell(\boldsymbol{x})$ is a degree 2ℓ polynomial and therefore satisfies the hypercontractivity property. For any $m \ge 1$, there exists $C > 0$ such that

$$\mathbb{E}_{\boldsymbol{x}}[F_\ell(\boldsymbol{x})^{2m}]^{1/(2m)} \le C \cdot \mathbb{E}_{\boldsymbol{x}}[F_\ell(\boldsymbol{x})^2]^{1/2}. \tag{80}$$

Let us bound the right hand side. We have

$$\mathbb{E}[F_\ell(\boldsymbol{x})^2] = \frac{1}{d^2 B(\mathscr{Q}^q;\ell)^2} \sum_{S,S'\in\mathcal{C}_\ell} r(S)r(S') \sum_{i,j,i',j'\in[d]} \omega(B_1,B_2,B_3,B_4),$$

where $B_1 = i+S$, $B_2 = j+S$, $B_3 = i'+S'$ and $B_4 = j'+S'$, and we denoted

$$\omega(B_1,B_2,B_3,B_4) = \mathbb{E}_{\boldsymbol{x}}\Big[Y_{B_1}(\boldsymbol{x})Y_{B_2}(\boldsymbol{x})Y_{B_3}(\boldsymbol{x})Y_{B_4}(\boldsymbol{x})\Big]\mathbb{1}_{B_1\neq B_2}\mathbb{1}_{B_3\neq B_4}.$$

Notice that $\omega(B_1,B_2,B_3,B_4) = 1$ if $B_1\Delta B_2 = B_3\Delta B_4$ (the symmetric difference) and 0 otherwise. In other words, every elements in $B_1\cup B_2\cup B_3\cup B_4$ appears exactly in 2 or 4 of these sets.

Let us fix $i\in[d]$ and $S\in\mathcal{C}_\ell$, and bound

$$\sum_{S'\in\mathcal{C}_{q,\ell}} r(S') \sum_{j,i'\neq j'\in[d]} \omega(B_1,B_2,B_3,B_4). \tag{81}$$

Denote $|B_1\Delta B_2| = 2k$ with $1\le k\le\ell$. In order for $\omega(B_1,B_2,B_3,B_4) = 1$, B_3 must contain exactly k points in $B_1\Delta B_2$ while B_4 must contain the remaining k points.

- **Case $k<\ell$.** There are at most ℓ^2 ways of choosing j such that $B_1\cap B_2\neq\emptyset$. Fixing j (i.e., B_1 and B_2) and S', then there are $2k\ell$ ways of choosing i' and $2k\ell$ ways of choosing j' such that $B_3\cap(B_1\Delta B_2)\neq\emptyset$ and $B_4\cap(B_1\Delta B_2)\neq\emptyset$. Hence the contribution of these terms in Eq. (81) is upper bounded by

$$\sum_{S'\in\mathcal{C}_\ell} r(S') \sum_{k=1}^{\ell-1} \ell^2\cdot(2k\ell)^2 \le 4\ell^7 \sum_{S'\in\mathcal{C}_\ell} r(S') = 4\ell^7 B(\mathscr{Q}^q;\ell). \tag{82}$$

- **Case $k=\ell$.** There are at most d ways of choosing j. Furthermore, for j fixed, there are at most $\binom{2\ell}{\ell}$ ways of choosing B_3 and B_4 such that $B_3\cup B_4 = B_1\cup B_2$ (note that $B_1\cap B_2=\emptyset$ and therefore $B_3\cap B_4=\emptyset$). Hence the contribution of these terms in Eq. (81) is upper bounded by

$$\sum_{S'\in\mathcal{C}_\ell, i',j'\in[d]} r(S')\cdot d\cdot \mathbb{1}_{B_3\cup B_4 = B_1\cup B_2} \le dq\binom{2\ell}{\ell}, \tag{83}$$

  where we used that $r(S')\le q$.

Combining Eqs. (82) and (83) and using there are $dB(\mathscr{Q}^q;\ell)$ choices for i and S_1, we get

$$\mathbb{E}[F_\ell(\boldsymbol{x})^2] \le \frac{1}{d^2 B(\mathscr{Q}^q;\ell)^2} \sum_{i\in[d], S\in\mathcal{C}_\ell} r(S)\Big[4\ell^7 B(\mathscr{Q}^q;\ell) + dq\binom{2\ell}{\ell}\Big]$$
$$= O_q(1)\cdot[d^{-1} + qB(\mathscr{Q}^q;\ell)^{-1}] = O_q(q^{-1}),$$

where we used that $\ell \ge 2$ and $B(\mathscr{Q}^q;\ell) = \Omega_q(q^\ell)$.

Using Eq. (80), we deduce

$$\mathbb{E}\Big[\max_{i\in[n]} |F_\ell(\boldsymbol{x}_i)|\Big] \le \mathbb{E}\Big[\max_{i\in[n]} F_\ell(\boldsymbol{x}_i)^{2m}\Big]^{1/(2m)} \le n^{1/(2m)}\mathbb{E}\Big[F_\ell(\boldsymbol{x}_i)^{2m}\Big]^{1/(2m)}$$
$$\le Cn^{1/(2m)}\mathbb{E}[F_\ell(\boldsymbol{x})^2]^{1/2} = n^{1/m}\cdot O_q(q^{-1/2}).$$

Using Markov's inequality and taking m sufficiently small yield Eq. (77).

Step 2. Bounding $\max_{i\in[n]}\Big|\Xi_\ell^{(d)}(\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}}[\Xi_\ell^{(d)}(\boldsymbol{x})]\Big|$**.**

The second bound (78) is obtained very similarly. Define $G_\ell : \mathscr{Q}^d \to \mathbb{R}$ to be

$$G_\ell(\boldsymbol{x}) = \Xi_\ell^{(d)}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}}[\Xi_\ell^{(d)}(\boldsymbol{x})] = \frac{1}{dR_\ell}\sum_{S\in\mathcal{C}_\ell} r(S)^2 \sum_{i\ne j\in[d]} Y_{i+S}(\boldsymbol{x})Y_{j+S}(\boldsymbol{x}). \tag{84}$$

Then, we have

$$\mathbb{E}[G_\ell(\boldsymbol{x})^2] = \frac{1}{d^2R_\ell^2}\sum_{S,S'\in\mathcal{C}_\ell} r(S)^2 r(S')^2 \sum_{i,i',j,j'\in[d]} \omega(B_1,B_2,B_3,B_4).$$

Further notice that following the same computation as in Eq. (69), we get

$$R_\ell = \sum_{S\in\mathcal{C}_\ell} r(S)^2 = \sum_{h=\ell}^{q}(q+1-h)^2\binom{h-2}{\ell-2} = \Omega_q(1)\cdot q^{1+\ell}.$$

Hence, the same computation as for F_ℓ in step 1 yields

$$\mathbb{E}[G_\ell(\boldsymbol{x})^2] \le \frac{1}{d^2R_\ell^2}\sum_{i\in[d], S\in\mathcal{C}_\ell} r(S)^2\Big[4\ell^7 R_\ell + dq^2\binom{2\ell}{\ell}\Big]$$
$$= O_q(1)\cdot[d^{-1} + q^2R_\ell^{-1}] = O_q(q^{-1}),$$

where we used that $\ell \ge 2$. We deduce Eq. (78) similarly to step 1. □

Lemma 3 (Hanson-Wright inequality)**.** *There exists a universal constant $c > 0$, such that for any $t > 0$ and $q^{1/\delta} \ge d \ge q \in \mathbb{N}$ for some $\delta > 0$, when $\boldsymbol{x} \in \mathrm{Unif}(\mathscr{Q}^d)$, we have*

$$\mathbb{P}\left(\sup_{k\ne l\in[d]} |\langle \boldsymbol{x}_{(k)}, \boldsymbol{x}_{(l)}\rangle|/q > t\right) \le 2q^{2/\delta}\exp\{-cq\cdot\min(t^2,t)\},$$

where we recall that $\boldsymbol{x}_{(k)} = (x_k,\dots,x_{k+q-1})$.

Proof of Lemma 3. For any $k \ne l$, denote $\boldsymbol{A} = (a_{ij})_{i,j\in[d]}$ the matrix with $a_{(k+i),(l+i)} = 1$ for $i = 0,\dots,q-1$ and $a_{ij} = 0$ otherwise, such that $\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle = \langle \boldsymbol{x}_{(k)}, \boldsymbol{x}_{(l)}\rangle$. Note that we have $\|\boldsymbol{A}\|_F = \sqrt{q}$, $\|\boldsymbol{A}\|_{\mathrm{op}} \le 1$ and $\mathbb{E}[\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle] = 0$. By Hanson-Wright inequality of vectors with independent sub-Gaussian entries (for example, see Theorem 1.1 in [42]), we have

$$\mathbb{P}\left(|\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle|/q > t\right) \le 2\exp\{-cq\cdot\min(t^2,t)\}.$$

Taking the union bound over $k \ne l$ concludes the proof. □

D Technical background of function spaces on the hypercube

Fourier analysis on the hypercube is a well studied subject [41]. The purpose of this section is to introduce some notations and objects that are useful in the statement and proofs in the main text.

D.1 Fourier basis

Denote $\mathscr{Q}^d = \{-1,+1\}^d$ the hypercube in d dimension, and τ_d to the uniform probability measure on $\mathscr{Q}^d$. All the functions will be assumed to be elements of $L^2(\mathscr{Q}^d, \tau_d)$ (which contains all the bounded functions $f : \mathscr{Q}^d \to \mathbb{R}$), with scalar product and norm denoted as $\langle \cdot, \cdot \rangle_{L^2}$ and $\| \cdot \|_{L^2}$:

$$\langle f, g \rangle_{L^2} \equiv \int_{\mathscr{Q}^d} f(\boldsymbol{x}) g(\boldsymbol{x}) \tau_d(\mathrm{d}\boldsymbol{x}) = \frac{1}{2^n} \sum_{\boldsymbol{x} \in \mathscr{Q}^d} f(\boldsymbol{x}) g(\boldsymbol{x}).$$

Notice that $L^2(\mathscr{Q}^d, \tau_d)$ is a 2^n dimensional linear space. By analogy with the spherical case we decompose $L^2(\mathscr{Q}^d, \tau_d)$ as a direct sum of $d+1$ linear spaces obtained from polynomials of degree $\ell = 0, \dots, d$

$$L^2(\mathscr{Q}^d, \tau_d) = \bigoplus_{\ell=0}^{d} V_{d,\ell}.$$

For each $\ell \in \{0, \dots, d\}$, consider the Fourier basis $\{Y^{(d)}_{\ell,S}\}_{S \subseteq [d], |S|=\ell}$ of degree ℓ, where for a set $S \subseteq [d]$, the basis is given by

$$Y^{(d)}_{\ell,S}(\boldsymbol{x}) \equiv x^S \equiv \prod_{i \in S} x_i.$$

It is easy to verify that (notice that $x_i^k = x_i$ if k is odd and $x_i^k = 1$ if k is even)

$$\langle Y^{(d)}_{\ell,S}, Y^{(d)}_{k,S'} \rangle_{L^2} = \mathbb{E}[x^S \times x^{S'}] = \delta_{\ell,k} \delta_{S,S'}.$$

Hence $\{Y^{(d)}_{\ell,S}\}_{S \subseteq [d], |S|=\ell}$ form an orthonormal basis of $V_{d,\ell}$ and

$$\dim(V_{d,\ell}) = B(\mathscr{Q}^d; \ell) = \binom{d}{\ell}.$$

We will omit the superscript (d) in $Y^{(d)}_{\ell,S}$ when clear from the context and write $Y_S := Y^{(d)}_{\ell,S}$.

We denote by P_ℓ the orthogonal projections to $V_{d,\ell}$ in $L^2(\mathscr{Q}^d)$. This can be written in terms of the Fourier basis as

$$\mathsf{P}_\ell f(\boldsymbol{x}) \equiv \sum_{S \subseteq [d], |S|=\ell} \langle f, Y_S \rangle_{L^2} Y_S(\boldsymbol{x}). \tag{85}$$

We also define $\mathsf{P}_{\le \ell} \equiv \sum_{k=0}^{\ell} \mathsf{P}_k$, $\mathsf{P}_{>\ell} \equiv \mathbf{I} - \mathsf{P}_{\le \ell} = \sum_{k=\ell+1}^{\infty} \mathsf{P}_k$, and $\mathsf{P}_{<\ell} \equiv \mathsf{P}_{\le \ell-1}$, $\mathsf{P}_{\ge \ell} \equiv \mathsf{P}_{>\ell-1}$.

D.2 Hypercubic Gegenbauer

We consider the following family of polynomials $\{Q^{(d)}_\ell\}_{\ell=0,\dots,d}$ that we will call *hypercubic Gegenbauer*, or G*egenbauer on the* d*-dimensional hypercube*, defined as

$$Q^{(d)}_\ell(\langle \boldsymbol{x}, \boldsymbol{y} \rangle) = \frac{1}{B(\mathscr{Q}^d; \ell)} \sum_{S \subseteq [d], |S|=\ell} Y^{(d)}_{\ell,S}(\boldsymbol{x}) Y^{(d)}_{\ell,S}(\boldsymbol{y}). \tag{86}$$

Notice that the right hand side only depends on $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ and therefore these polynomials are well defined. In particular,

$$\langle Q^{(d)}_\ell(\langle \mathbf{1}, \cdot \rangle), Q^{(d)}_k(\langle \mathbf{1}, \cdot \rangle) \rangle_{L^2} = \frac{1}{B(\mathscr{Q}^d; k)} \delta_{\ell k}.$$

Hence $\{Q^{(d)}_\ell\}_{\ell=0,\dots,d}$ form an orthogonal basis of $L^2(\{-d, -d+2, \dots, d-2, d\}, \tilde{\tau}^1_d)$ where $\tilde{\tau}^1_d$ is the distribution of $\langle \mathbf{1}, \boldsymbol{x} \rangle$ when $\boldsymbol{x} \sim \tau_d$, i.e., $\tilde{\tau}^1_d \sim 2\mathrm{Bin}(d, 1/2) - d/2$.

It is easy to check more generally that

$$\langle Q^{(d)}_\ell(\langle \boldsymbol{x}, \cdot \rangle), Q^{(d)}_k(\langle \boldsymbol{y}, \cdot \rangle) \rangle_{L^2} = \frac{1}{B(\mathscr{Q}^d; k)} Q_k(\langle \boldsymbol{x}, \boldsymbol{y} \rangle) \delta_{\ell k}.$$

Furthermore, Eq. (86) imply that —up to a constant— $Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle)$ is a representation of the projector onto the subspace of degree-k polynomials

$$(\mathsf{P}_k f)(\boldsymbol{x}) = B(\mathscr{Q}^d; k) \int_{\mathscr{Q}^d} Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle)\, f(\boldsymbol{y})\, \tau_d(\mathrm{d}\boldsymbol{y})\,. \tag{87}$$

For a function $\sigma(\cdot/\sqrt{d}) \in L^2(\{-d, -d+2, \dots, d-2, d\}, \tilde{\tau}_d^1)$, denote its hypercubic Gegenbauer coefficients $\xi_{d,k}(\sigma)$ to be

$$\xi_{d,k}(\sigma) = \int_{\{-d,-d+2,\dots,d-2,d\}} \sigma(x/\sqrt{d}) Q_k^{(d)}(x) \tilde{\tau}_d^1(\mathrm{d}x). \tag{88}$$

To any inner-product kernel $H_d(\boldsymbol{x}_1, \boldsymbol{x}_2) = h_d(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle/d)$, with $h_d(\,\cdot\,/\sqrt{d}) \in L^2(\{-d, -d+2, \dots, d-2, d\}, \tilde{\tau}_d^1)$, we can associate a self adjoint operator $\mathscr{H}_d : L^2(\mathscr{Q}^d) \to L^2(\mathscr{Q}^d)$ via

$$\mathscr{H}_d f(\boldsymbol{x}) \equiv \int_{\mathscr{Q}_d} h_d(\langle \boldsymbol{x}, \boldsymbol{x}_1\rangle/d)\, f(\boldsymbol{x}_1)\, \tau_d(\mathrm{d}\boldsymbol{x}_1)\,. \tag{89}$$

By permutation invariance, the space V_k of homogeneous polynomials of degree k is an eigenspace of $\mathscr{H}_d$, and we will denote the corresponding eigenvalue by $\xi_{d,k}(h_d)$. In other words $\mathscr{H}_d f(\boldsymbol{x}) \equiv \sum_{k=0}^{q} \xi_{d,k}(h_d) \mathsf{P}_k f$. The eigenvalues can be computed via

$$\xi_{d,k}(h_d) = \int_{\{-d,-d+2,\dots,d-2,d\}} h_d\big(x/d\big) Q_k^{(d)}(x) \tilde{\tau}_d^1(\mathrm{d}x)\,. \tag{90}$$

D.3 Hermite polynomials

The Hermite polynomials $\{\mathrm{He}_k\}_{k\ge 0}$ form an orthogonal basis of $L^2(\mathbb{R}, \gamma)$, where $\gamma(\mathrm{d}x) = e^{-x^2/2}\mathrm{d}x/\sqrt{2\pi}$ is the standard Gaussian measure, and He_k has degree k. We will follow the classical normalization (here and below, expectation is with respect to $G \sim \mathsf{N}(0,1)$):

$$\mathbb{E}\big\{\mathrm{He}_j(G)\, \mathrm{He}_k(G)\big\} = k!\, \delta_{jk}\,. \tag{91}$$

As a consequence, for any function $g \in L^2(\mathbb{R}, \gamma)$, we have the decomposition

$$g(x) = \sum_{k=0}^{\infty} \frac{\mu_k(g)}{k!}\, \mathrm{He}_k(x)\,, \qquad \mu_k(g) \equiv \mathbb{E}\big\{g(G)\, \mathrm{He}_k(G)\big\}\,. \tag{92}$$

The Hermite polynomials can be obtained as high-dimensional limits of the Gegenbauer polynomials introduced in the previous section. Indeed, the Gegenbauer polynomials (up to a $\sqrt{d}$ scaling in domain) are constructed by Gram-Schmidt orthogonalization of the monomials $\{x^k\}_{k\ge 0}$ with respect to the measure $\tilde{\tau}_d^1$, while Hermite polynomial are obtained by Gram-Schmidt orthogonalization with respect to γ. Since $\tilde{\tau}_d^1 \Rightarrow \gamma$ (here $\Rightarrow$ denotes weak convergence), it is immediate to show that, for any fixed integer k,

$$\lim_{d\to\infty} \mathrm{Coeff}\{Q_k^{(d)}(\sqrt{d}x)\, B(\mathscr{Q}^d; k)^{1/2}\} = \mathrm{Coeff}\left\{\frac{1}{(k!)^{1/2}}\, \mathrm{He}_k(x)\right\}\,. \tag{93}$$

Here and below, for P a polynomial, $\mathrm{Coeff}\{P(x)\}$ is the vector of the coefficients of P. As a consequence, for any fixed integer k, we have

$$\mu_k(\sigma) = \lim_{d\to\infty} \xi_{d,k}(\sigma)(B(\mathscr{Q}^d; k)k!)^{1/2}, \tag{94}$$

where $\mu_k(\sigma)$ and $\xi_{d,k}(\sigma)$ are given in Eq. (92) and (88).

D.4 Hypercontractivity of uniform distributions on the hypercube

By Holder's inequality, we have $\|f\|_{L^p} \le \|f\|_{L^q}$ for any f and any $p \le q$. The reverse inequality does not hold in general, even up to a constant. However, for some measures, the reverse inequality will hold for some sufficiently nice functions. These measures satisfy the celebrated hypercontractivity properties [4, 5, 8, 25].

Lemma 4 (Hypercube hypercontractivity [4])**.** *For any $\ell = \{0, \dots, d\}$ and $f_d \in L^2(\mathscr{Q}^d)$ to be a degree ℓ polynomial, then for any integer $q \ge 2$, we have*

$$\|f_d\|^2_{L^q(\mathscr{Q}^d)} \le (q-1)^\ell \cdot \|f_d\|^2_{L^2(\mathscr{Q}^d)}.$$