How Wide Convolutional Neural Networks Learn Hierarchical Tasks

Francesco Cagnetta * 
Institute of Physics 
École Polytechnique Fédérale de Lausanne 
francesco.cagnetta@epfl.ch

Alessandro Favero * 
Institute of Physics 
École Polytechnique Fédérale de Lausanne 
alessandro.favero@epfl.ch

Matthieu Wyart 
Institute of Physics 
École Polytechnique Fédérale de Lausanne 
matthieu.wyart@epfl.ch

Abstract

Despite their success, understanding how convolutional neural networks (CNNs) can efficiently learn high-dimensional functions remains a fundamental challenge. A popular belief is that these models harness the compositional and hierarchical structure of natural data such as images. Yet, we lack a quantitative understanding of how such structure affects performances, e.g. the rate of decay of the generalisation error with the number of training samples. In this paper we study deep CNNs in the kernel regime: i) we show that the spectrum of the corresponding kernel and its asymptotics inherit the hierarchical structure of the network; ii) we use generalisation bounds to prove that deep CNNs adapt to the spatial scale of the target function; iii) we illustrate this result by computing the rate of decay of the error in a teacher-student setting, where a deep CNN is trained on the output of another deep CNN with randomly-initialised parameters. We find that if the teacher function depends on certain low-dimensional subsets of the input variables, then the rate is controlled by the effective dimensionality of these subsets. Conversely, if the teacher function depends on the full set of input variables, then the error rate is inversely proportional to the input dimension. Interestingly, this implies that despite their hierarchical structure, the functions generated by deep CNNs are too rich to be efficiently learnable in high dimension.

1 Introduction

Deep convolutional neural networks (CNNs) are particularly successful in certain tasks such as image classification. Such tasks generally entail the approximation of functions of a large number of variables, for instance the number of pixels which determine the content of an image. Learning a generic high-dimensional function is plagued by the curse of dimensionality: the rate at which the generalisation error $\epsilon$ decays with the number of training samples $n$ vanishes as the dimensionality $d$ of the input space grows, i.e. $\epsilon(n) \sim n^{-\beta}$ with $\beta = O(1/d)$ [1]. Therefore, the success of CNNs in classifying data whose dimension can be in the hundreds or more [2,3] points to the existence of some underlying structure in the task that CNNs can leverage [4]. Understanding the structure of learnable tasks is arguably one of the most fundamental problems in deep learning, and also one of central practical importance—as it determines how many examples are required to learn up to a certain error. A popular hypothesis is that learnable tasks are compositional and hierarchical: features

*Equal contribution.

Preprint. Under review.
at any scale are made of sub-features of smaller scales. Although many works have investigated this hypothesis \[5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17\], there are no available predictions for the exponent \(\beta\) for deep CNNs trained on tasks with a truly hierarchical structure. In this paper, we perform such computation in the overparameterised regime where neural networks behave as kernel methods \[18, 19, 20\].

To model the structure of natural tasks, we introduce spatially-localised target functions that depend only on a subset of adjacent coordinates of the input \[14\]—the spatial scale of the target. Then, we prove that simple deep CNNs (such as the one illustrated in Fig. 1) are adaptive to the spatial scale, in the sense that they achieve performances comparable to that of an architecture a priori fine-tuned to the spatial scale of the target. In addition, we can make the task hierarchical by using the outputs of deep CNNs as target functions; this addition does not improve performance, implying that having such kind of hierarchical structure in the task is insufficient to beat the curse of dimensionality.

1.1 Our contributions

This paper studies the generalisation properties of deep CNNs with non-overlapping patches and no pooling (defined in Sec. 2, see Fig. 1 for an illustration) trained by empirical minimisation of the mean squared loss on a target function \(f^*\). In particular, we consider the infinite-width limit (Sec. 3), where generalisation is fully characterised by the spectrum of the integral operator of the Neural Tangent Kernel (NTK) of the network \[18, 19, 20, 21\]. In simple terms, the projections on the eigenfunctions with larger eigenvalues can be learnt (up to a fixed generalisation error) with fewer training points.

Spectral bias of deep hierarchical CNNs in the kernel regime. Due to the network architecture, the hidden neurons of each layer depend only on a subset of the input variables, known as the receptive field of that neuron (highlighted by coloured boxes in Fig. 1, left panel). We find that the eigenfunctions of the NTK of a hierarchical CNN of depth \(L + 1\) can be organised into sectors \(l \in \{1, \ldots, L\}\) associated with the hidden layers of the network (Sec. 3). In particular, the eigenfunctions of each sector depend only on the receptive fields of the neurons of the corresponding hidden layer. We denote with \(d_{\text{eff}}(l)\) the size of the receptive fields of neurons in the \(l\)-th hidden layer, so that the eigenfunctions of the \(l\)-th sector are effectively functions of \(d_{\text{eff}}(l)\) variables.

Adaptivity to the spatial structure of the target. We use the above result to prove that deep CNNs can adapt to the spatial scale of the target function (Sec. 4). More specifically, by using rigorous bounds from the theory of kernel ridge regression \[22\] (reviewed in the first paragraph of Sec. 4), we show that learning a spatially-localised target \(f^*\) which only depends on \(d_{\text{eff}}\) adjacent coordinates of the \(d\)-dimensional input with the kernel of a CNN and optimal regularisation leads to an exponent \(\beta \geq O(1/d_{\text{eff}})\) (Cor. 4.1, see Fig. 1 for a pictorial representation). We find a similar picture in ridgeless regression by using results derived with the replica method \[23, 24\] (Sec. 5). These rates are much closer to the Bayes-optimal result—realised when the architecture is fine-tuned to the structure of the target—than \(\beta = O(1/d)\) obtained with the kernel of a fully-connected network. We confirm these results through extensive numerical studies.

1.2 Related work

The benefits of shallow CNNs in the kernel regime have been investigated in \[13, 14, 15, 4, 12, 14\], and later \[15, 18\], studied generalisation properties of shallow CNNs, finding that they are able to beat the curse of dimensionality on local target functions. However, these architectures can only approximate functions of single patches of the input or linear combinations thereof. \[13\], in addition, includes generic pooling layers and begins considering the role of depth by studying the approximation properties of kernels which are integer powers of other kernels. We generalise this line of work by studying CNNs of any depth with nonanalytic (ReLU) activations: we find that the depth and nonanalyticity of the resulting kernel are crucial for understanding the inductive bias of deep CNNs. This result should also be contrasted with the spectrum of the kernels of deep fully-connected networks, whose
asymptotics do not depend on depth \cite{25}. Furthermore, we extend the analysis of generalisation to target functions that have a hierarchical structure similar to that of the networks themselves.

\cite{17} derives bounds on the spectrum of the kernels of deep CNNs. However, they consider only filters of size one in the first layer and do not include a theoretical analysis of generalisation. Instead, we allow filters of general dimension and give tight estimates of the asymptotic behaviour of eigenvalues, which allow us to predict generalisation properties. \cite{16} is the closest to our work, as it also investigates the spectral bias of deep CNNs in the kernel regime. However, it considers a different limit where both the input dimension and the number of training points diverge and does not characterise the asymptotic decay of generalisation error with the number of training samples.

2 Notation and setup

Our work considers CNNs with nonoverlapping patches and no pooling layers: these networks are fully characterised by the depth $L + 1$ (or number of hidden layers $L$) and a set of filters sizes $\{s_i\}_i$ (one per hidden layer). We call such networks hierarchical CNNs.

**Definition 2.1 (L-hidden-layers hierarchical CNN)** Denote by $\sigma$ the normalised ReLU function, $\sigma(x) = \sqrt{2} \max(0, x)$. For each input $x \in \mathbb{R}^d$ and $s$ a divisor of $d$, denote by $x_i$ the $i$-th $s$-dimensional patch of $x$, $x_i = (x_{i-1 \times s+1}, \ldots, x_{i \times s})$ for all $i = 1, \ldots, d/s$. The output of a L-

\footnote{Notice that all our results can be readily extended to image-like input signals $\{x_{ij}\}_{i,j}$ or tensorial objects with an arbitrary number of indices.}
hidden-layers hierarchical neural network can be defined recursively as follows.

\[
\begin{align*}
    f^{(l)}_{h,i}(x) &= \sigma \left( w^{(l)}_h \cdot x_i \right), \quad \forall h \in [1 \ldots H^l], \forall i \in [1 \ldots p^l]; \\
    f^{(l)}_{h,i}(x) &= \sigma \left( \frac{1}{\sqrt{H_{l-1}}} \sum_{h'} w^{(l)}_{h,h'} \cdot \left( f^{(l-1)}_{h',i} \right) \right), \quad \forall h \in [1 \ldots H^l], i \in [1 \ldots p^l], l \in [2 \ldots L]; \\
    f(x) &= f^{(L+1)}(x) = \frac{1}{\sqrt{H_L}} \sum_{h=1}^{H_L} \sum_{i=1}^{p_L} w^{(L+1)}_{h,i} f^{(L)}_{h,i}(x).
\end{align*}
\]

\(H_l\) denotes the width of the \(l\)-th layer, \(s_l\) the filter size (\(s_1 = s\)), \(p_l\) the number of patches (\(p_1 \equiv p = d/s\)), \(w^{(l)}_{h,i} \in \mathbb{R}^{s_l}, \quad w^{(l+1)}_{h,i} \in \mathbb{R}^{s_l}, \quad w^{(L+1)}_{h,i} \in \mathbb{R}\).

Hierarchical CNNs are best visualised by considering the Directed Acyclic Graph (DAG) obtained by setting \(H_l = 1 \forall l\), as in Fig. 1. With nonoverlapping patches, the DAG is an ordered tree, whose root is the output (empty circle at the top of the figure) and the leaves are the input coordinates (squares at the bottom). All the other nodes represent neurons. The height of a node is the distance between that node and the input nodes: all neurons belonging to the same hidden layer have equal height. The tree structure highlights that the post-activations \(f^l\) of the \(l\)-th layer depend only on a subset of the input variables, also known as the receptive field.

Since the first layer of a hierarchical CNN acts on \(s_1\)-dimensional patches of the input, it is convenient to consider each \(d\)-dimensional input signal as the concatenation of \(p\) \(s\)-dimensional patches, with \(s = s_1\) and \(p \times s = d\). We assume that each patch is normalised to 1, so that the input space is a product of \(p\) \(s\)-dimensional unit spheres (called multisphere in [17]):

\[
    M^p S^{s-1} := \prod_{i=1}^{p} S^{s-1} \subset \mathbb{S}^{d-1}.
\]

Notice that the \(s\)-dimensional patches are also the receptive fields of the first-hidden-layer neurons (as in the blue rectangle in Fig. 1 for \(s = 3\)). In general, the receptive field of a neuron in the \(l\)-th hidden layer with \(l > 1\) is a group of \(\prod_{l'=3}^{l} \mathbb{S}^{s_{l'}}\) adjacent patches (as in the orange rectangle of Fig. 1 for \(l = 2, s_2 = 2\) or the green rectangle for \(l = 3, s_3 = s_2 = 2\)), which we refer to as a meta-patch. Due to the correspondence between the receptive fields, each meta-patch is identified with one path on the network’s DAG: the path which connects the output node to the hidden neuron whose receptive field coincides with the meta-patch. If such hidden neuron belongs to the \(l\)-th hidden layer, the path is specified by a tuple of \(L - l + 1\) indices, \(i_{l+1} \ldots i_{L+1} = i_{l+1} \ldots i_{L+1}\), where each index indicates which branch to select when descending from the root of the DAG to the desired node. With this notation, \(x_{i_{l+1} \ldots i_{L+1}}\) denotes one of the \(p_l\) meta-patches of size \(\prod_{l'=3}^{L} \mathbb{S}^{s_{l'}}\). Because of the normalisation of the \(s_1\)-dimensional patches, each meta-patch has an effective dimensionality which is lower than its size:

\[
    x_{i_{2 \ldots L+1}} \in \mathbb{S}^{s_1-1} \Rightarrow \begin{cases} 
    d_{\text{eff}}(1) := \dim(x_{i_2 \ldots L+1}) = (s_1 - 1), \\
    d_{\text{eff}}(l) := \dim(x_{i_{l+1} \ldots i_{L+1}}) = (s_1 - 1) \prod_{l'=2}^{L} \mathbb{S}^{s_{l'}}, \quad \forall l \in [2 \ldots L].
\end{cases}
\]

### 3 Hierarchical kernels and their spectra

In order to study the generalisation properties of the hierarchical models of Def. 2.1 we consider the infinite-width limit \(H_l \to \infty\). Because of the aforementioned equivalence with kernel methods, this limit allows us to deduce the generalisation properties of the network from the spectrum of a kernel. We consider, in particular, two kernels: the Neural Tangent Kernel (NTK), corresponding to training all the network parameters [18], and the Random Feature Kernel (RFK), corresponding to training only the weights of the linear output layer [26] [27]. In both cases, the kernel reads:

\[
    K(x, y) = \sum_{\text{trained params } \theta} \partial_0 f(x) \partial_0 f(y).
\]

The NTK and RFK of deep CNNs have been derived previously [28]. For hierarchical CNNs one has:
Definition 3.1 (Kernels of hierarchical CNNs) Let $x, y \in \mathbb{M}^{p \mathbb{S}^{s-1}}$. Denote tuples of the kind $t_{i_{t-1} \ldots i_{i_{m}}} \mid t_{i_{t-1}}$ with $i_{t\rightarrow m}$ for $m \geq l$. For $m < l$, let $t_{l\rightarrow m}$ denote the empty tuple. For each tuple $t_{2\rightarrow L+1}$, denote with $t_{2\rightarrow L+1}$ the scalar product between the $s$-dimensional patches of $x$ and $y$ identified by the same tuple, i.e.

$$t_{2\rightarrow L+1} = x_{2\rightarrow L+1} \cdot y_{2\rightarrow L+1}$$

(5)

For $1 \leq l \leq L + 1$, denote with $\{t_{2\rightarrow L+1}\}_{t_{2\rightarrow L+1}}$ the sequence of $t$’s obtained by letting the indices of the tuple $t_{2\rightarrow l}$ vary in their respective range. Consider a hierarchical CNN with $L$ hidden layers, filter sizes $(s_1, \ldots, s_L)$, $p_1 \geq 1$ and all the weights $w_{1, i}, w_{1, j}, (w_{L+1})$ initialised as Gaussian random numbers with zero mean and unit variance. Therefore, the large-

$$t_{2\rightarrow L+1}$$

For the sake of clarity, we limit the discussion in the main paper to the case of the same hierarchical CNN has the same structure as the RFK, but the function used in the recursion is different. Instead of $K_{RFK}^{(l)}(t_{2\rightarrow L+1}) = \kappa_1(t_{2\rightarrow L+1})$, one has $K_{RFK}^{(l)} = \kappa_1(\sum K_{RFK}^{(l-1)})$, $\forall l \in [2 \ldots L]$ if $L > 1$;

$$K_{RFK}^{(l)}(t_{2\rightarrow L+1}) = 1 \frac{1}{s_1} \sum t_{i_{t\rightarrow L+1}} = \frac{1}{p_L} \sum K_{RFK}^{(L)}(t_{2\rightarrow L+1})$$

(6)

The NTK of the same hierarchical CNN has the same structure as the RFK, but the function used in the recursion is different. Instead of $K_{NTK}^{(l)} = \kappa_1(\sum K_{RFK}^{(l-1)})$, $\forall l \in [2 \ldots L]$ if $L > 1$;

$$K_{NTK}^{(l)} = \kappa_1(\sum K_{RFK}^{(l-1)})$$

(7)

with $dp(x)$ denoting the uniform distribution of input points on the multispheres. The spectrum of this operator provides, via Mercer’s theorem [29], an alternative representation of the kernel $K(x, y)$ and a basis for the space of functions that the kernel can approximate. The asymptotic decay of the eigenvalues, in particular, is crucial for studying the generalisation properties of the kernel. Since the input space is a product of $s$-dimensional unit spheres, and the kernel is a function of the $p$ scalar products between corresponding $s$-dimensional patches of $x$ and $y$, the eigenfunctions of $T_K$ are products of spherical harmonics acting on the patches (see App. A for the relevant background). For the sake of clarity, we limit the discussion in the main paper to the case $s = 2$, where, since each patch $x_t$ is entirely determined by an angle $\theta_t$, the multisphere $\mathbb{M}^{p \mathbb{S}^{s-1}}$ reduces to the $p$-dimensional torus and the eigenfunctions to $p$-dimensional plane waves; with $\theta := (\theta_1, \ldots, \theta_p)$, $\mathbb{Y}_k(\theta) \propto e^{ikT}$, which is labelled with $k := (k_1, \ldots, k_p)$. The general case is presented in Appendix. After setting $t_{2\rightarrow L+1} = \cos(\theta_{2\rightarrow L+1})$, in the kernel, the eigenvalues can be computed via the $p$-dimensional Fourier transform. Therefore, the large-$k$ asymptotics of the eigenvalues are controlled by the nonanaticities of the kernel as a function of $\theta$ [30]. In particular, kernels $K_{(t_{2\rightarrow L+1})_{2\rightarrow L+1}}$ such as those of Def. 3.1 have nonanaticities at $t_{2\rightarrow L+1} = \pm 1$, corresponding to $\theta_{2\rightarrow L+1} = 0, \pm \pi$. The singular expansion around the nonanaticities is influenced by the hierarchical structure. For
instance, since \( \kappa_1(\cos(\theta)) = 1 - \theta^2/2 + c|\theta|^3 + O(|\theta|^5) \) for \( \theta \to 0 \), then

\[
K_{\text{RFK}}^{(1)}(\cos(\theta_{i_2 \to L+1})) = 1 - \theta_{i_2 \to L+1}^2/2 + c|\theta_{i_2 \to L+1}|^3 + O(|\theta_{i_2 \to L+1}|^5),
\]

\[
K_{\text{RFK}}^{(2)} \left( \cos(\theta_{i_2 \to L+1}) \right)_{i_2} = \kappa_1 \left( 1 - \frac{1}{s_2} \sum_{i_2} \theta_{i_2 \to L+1}^2/2 + \frac{c}{s_2} \sum_{i_2} |\theta_{i_2 \to L+1}|^3 + O(|\theta_{i_2 \to L+1}|^5) \right)
\]

\[
= 1 - \frac{1}{s_2} \sum_{i_2} \theta_{i_2 \to L+1}^2/2 + \frac{c}{s_2} \sum_{i_2} |\theta_{i_2 \to L+1}|^3 + c'' \left( \frac{1}{s_2} \sum_{i_2} \theta_{i_2 \to L+1}^2 \right)^{3/2} + O(\theta_{i_2 \to L+1}^5),
\]

and so on. After only two steps, one already notices that the singular expansion at layer \( l \) consists of all the terms in the expansion of the previous layer averaged over \( i_l \), plus a new nonanalytic term of the kind \( c(\sum_{i_l} \theta^2)^{3/2} \). As a result, the singular expansion of \( K^{(l+1)} \) contains one nonanalytic term per hidden layer, each depending on the norm of the \( \theta \)'s within a certain meta-patch. Each of these singular terms turns out to produce eigenvalues with a different dependency on \( k \): this result, detailed in the following theorem, is proven in Appendix C.

**Theorem 3.1 (Spectrum of hierarchical kernels)** Let \( T_K \) be the integral operator associated with a \( d \)-dimensional hierarchical kernel of depth \( L + 1 \), \( L > 1 \), with filter sizes \( (s_1, \ldots, s_L) \). Eigenvalues and eigenfunctions of \( T_K \) can be organised into \( L \) sectors associated with the hidden layers of the kernel/network. For each \( 1 \leq l \leq L \), the \( l \)-th sector consists of \( \prod_{\nu=1}^{L-1} s_{\nu} \)-local eigenfunctions: functions of a single meta-patch \( x_{i_1 \to i_{L+1}} \) which cannot be written as linear combinations of functions of smaller meta-patches. The labels \( k \) of these eigenfunctions are such that there is a meta-patch \( k_{i_{l+1} \to i_{L+1}} \) of \( k \) with no vanishing sub-meta-patches and all the \( k_i \)'s outside of \( k_{i_{l+1} \to i_{L+1}} \) are 0 (because the eigenfunction is constant outside of \( x_{i_1 \to i_{L+1}} \)). The corresponding eigenvalue is degenerate with respect to the location of the meta-patch: we call it \( \Lambda^{(l)}_{k_{i_{l+1} \to i_{L+1}}} \). When \( \|k_{i_{l+1} \to i_{L+1}}\| \to \infty \), with \( k = \|k_{i_{l+1} \to i_{L+1}}\| \),

i. if \( s_1 = 2 \), then

\[
\Lambda^{(l)}_{k_{i_{l+1} \to i_{L+1}}} = C_{2,l} k^{-2\nu-\nu_{\text{eff}}(l)} + o \left( k^{-2\nu-\nu_{\text{eff}}(l)} \right),
\]

with \( \nu_{\text{NTK}} = 1/2 \), \( \nu_{\text{RFK}} = 3/2 \) and \( \nu_{\text{eff}} \) the effective dimensionality of the meta-patches defined in Eq. 3 \( C_{2,l} \) is a strictly positive constant for \( l \geq 2 \) whereas for \( l = 1 \) it can take two distinct strictly positive values depending on the parity of \( k_{i_2 \to L+1} \):

ii. if \( s_1 \geq 3 \), then for fixed non-zero angles \( k/k_k \),

\[
\Lambda^{(l)}_{k_{i_{l+1} \to i_{L+1}}} = C_{s_1,l} \left( \frac{k_{i_{l+1} \to i_{L+1}}}{k} \right) k^{-2\nu-\nu_{\text{eff}}(l)} + o \left( k^{-2\nu-\nu_{\text{eff}}(l)} \right),
\]

where \( C_{s_1,l} \) is a positive function for \( l \geq 2 \), whereas for \( l = 1 \) it is a strictly positive constant which depends on the parity of \( k_{i_2 \to L+1} \).

Some remarks are in order. First, if \( p_L = 1 \), the eigenfunctions of the \( L \)-th sector are supported on the full input space. Then, if \( \Lambda_k > 0 \) for all \( k \), hierarchical kernels are able to approximate any function on the multisphere. Secondly, the asymptotic scaling of an eigenvalue can be determined easily by looking at the corresponding eigenfunction and the tree-graph representation of the original network (see Fig. 1). Consider for instance an eigenfunction which depends only on two components of \( x \), \( x_i \) and \( x_j \), the sector to which such eigenfunction belongs coincides with the height of the lowest common ancestor of the \( i \)-th and \( j \)-th leaves. The effective dimensionality \( \nu_{\text{eff}}(l) \) of the sector is found by counting the total number of leaves connected to the lowest common ancestor, then subtracting the number of constraints coming from the patches’ normalisation. The eigenvalue scales with a negative power of the degree norm \( k \) obtained by summing \( \nu_{\text{eff}}(l) \) with a term \( 2\nu \) depending on the singular behaviour of the kernel. As a result, the decay of an eigenvalue with \( k \) is slower if the associated eigenfunction depends on a few adjacent patches. This is a property of kernels corresponding to hierarchical architectures which use nonlinear activation functions at all layers. Such a feature disappears if all hidden layers apart from the first have polynomial \( [13] \) or infinitely smooth \( [31] [32] \) activation functions or if the kernels are assumed to factorise over patches as in [17].
4 Generalisation properties and adaptivity to spatial structure

In this section, we study the generalisation properties of hierarchical NTKs and RFKs and prove a form of adaptivity to the spatial structure of the target function. We follow the classical analysis of [22] for kernel ridge regression and refer the reader to [33, 13] for a modern treatment.

General theory of kernel regression and source-capacity conditions. Given a set of \( n \) training points \( \{(x_\mu, y_\mu)\}_{\mu=1}^n \) for some probability density function \( p(x, y) \) and a regularisation parameter \( \lambda > 0 \), the kernel ridge regression estimate of the functional relation between \( x \)’s and \( y \)’s, or predictor, is

\[
    f^*_\lambda(x) = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{\mu=1}^n (f(x_\mu) - y_\mu)^2 + \lambda \|f\|_{\mathcal{H}} \right\},
\]

where \( \mathcal{H} \) is the Reproducing Kernel Hilbert Space (RKHS) of a (hierarchical) kernel \( \mathcal{K} \). If \( f(x, \theta) \) denotes the model from which the kernel was obtained via Eq. 4, the space \( \mathcal{H} \) is contained in the span of the network features \( \{\partial_\theta f(x, \theta)\}_{\theta} \) in the infinite-width limit [18, 26]. Alternatively, \( \mathcal{H} \) can be defined via the kernel’s eigenvalues \( \Lambda_k \) and eigenfunctions \( \hat{Y}_k \): denoting with \( f_k \) the projection of a function \( f \) onto the kernel eigenfunctions, then \( f \) belongs to \( \mathcal{H} \) if it belongs to the span of the eigenfunctions and

\[
    \|f^*\|_{\mathcal{H}}^2 = \sum_{k \geq 0} \Lambda_k^{-1} (f_k)^2 < +\infty. \tag{12}
\]

We quantify the generalisation properties of the kernel with the generalisation error and its expectation over training sets of fixed size \( n \) (denoted with \( E_n \))

\[
    \epsilon(f^*_\lambda) = \int dx dy p(x, y) (f^*_\lambda(x) - y)^2; \quad \tau(\lambda, n) = E_n[\epsilon(f^*_\lambda)], \tag{13}
\]

or the excess generalisation error, obtained by subtracting from \( \tau(\lambda, n) \) the error of the optimal predictor \( f^*(x) = \int dy p(x, y) y \). The decay of the error with \( n \) can be controlled via two exponents, depending on the details of the kernel and the target function. Specifically, if one can find \( \alpha \geq 1 \) and \( r \geq 1 - 1/\alpha \) satisfying the following conditions,

- capacity: \( \text{Tr} \left( \mathcal{T}_\mathcal{K}^{1/\alpha} \right) = \sum_{k \geq 0} (\Lambda_k)^{1/\alpha} < +\infty \),
- source: \( \left\| \mathcal{T}_\mathcal{K}^{-1/\alpha} f^* \right\|_{\mathcal{H}}^2 = \sum_{k \geq 0} (\Lambda_k)^{-r} (f_k^*)^2 < +\infty, \tag{14} \)

then, by choosing a \( n \)-dependent regularisation parameter \( \lambda_n \sim n^{-\alpha/(\alpha r + 1)} \), one gets the following bound on generalisation [22, 33]:

\[
    \tau(\lambda_n, n) - \epsilon(f^*) \leq C n^{-\frac{\alpha r}{\alpha r + 1}}. \tag{15}
\]

The capacity condition depends only on the kernel spectrum: \( \alpha \geq 1 \) since \( \text{Tr}(\mathcal{T}_\mathcal{K}) \) is finite [34]; if the ordered eigenvalues decay as a power \( > 1 \) of their rank (as it is the case when \( \Lambda_k \) scales as a power of \( k = \|k\| \)) then \( \alpha \) can be set to such power; any \( \alpha \) can be chosen if the eigenvalues decay faster than any power law. The source condition characterises the regularity of the target function relative to the kernel and can be used to prove that kernel methods are adaptive to the smoothness of the target function [33]: this is because the projections of smoother targets on the eigenfunctions display a faster decay with \( k \), thus allowing to choose a larger \( r \) satisfying the source condition and leading to better generalisation performances.

The following corollary of [Thm. 3.1] shows that, when the spectrum can be partitioned as in [Thm. 3.1], \( r \) also depends on the spatial scale of \( f^* \), thus showing that hierarchical kernels display adaptivity to targets which depend only on a subset of the input variables. Specific examples of bounds are considered explicitly in Sec. 5.

Corollary 4.1 (Adaptivity to spatial structure) Let \( x, y \in \mathbb{R}^{s-1} \) with \( s = 2 \) and \( \mathcal{K} \) be the RFK/NTK of a hierarchical CNN with \( L \) hidden layers, filter sizes \((s_1, \ldots, s_L)\) and \( p_L \geq 1 \). Denote with \( \Lambda_k \) the eigenvalues of the corresponding integral operator \( \mathcal{T}_\mathcal{K} \). Consider a target function \( f^* \)
on $M^n$ for $s = 1, \ldots, L$. If there is $l = 1, \ldots, L$ such that $f^*$ depends only on some meta-patch $x_{i_l+1 \ldots, L+1}$ of the input $x$, then only the first $l$ sectors of the spectrum of $T_K$ contribute to the source condition, i.e.

$$\left\| T_K^{\frac{1}{2}} f^* \right\|_H^2 = \sum_{l'=1}^l \sum_{k_{l'+1 \ldots, L+1}} \sum_{\ell_{l'+1 \ldots, L+1}} \left( \lambda_k^{(l')} \right)^{-r} \left( f_{k_{l'+1 \ldots, L+1}}^{\star} \right)^2 \left( f_{\ell_{l'+1 \ldots, L+1}}^{\star} \right)^2.$$ \hspace{1cm} \hspace{1cm} (16)

The same holds if $f^*$ is a linear combination of such functions.

**Application on target functions of controlled smoothness and $d_{\text{eff}}$.** The corollary follows immediately from the structure of the spectrum presented in Thm. 3.1 Its implications are again best understood in the case of two-dimensional patches $s = 2$. Fix $l \in [1 \ldots L]$ and consider a target function $f^*$ which only depends on the meta-patch $x_{i_l+1 \ldots, L+1}$. Combining the source condition in Eq. 10 with the asymptotic scaling of eigenvalues (Eq. 9), we get

$$\left\| T_K^{\frac{1}{2}} f^* \right\|_H^2 = +\infty \Leftrightarrow \sum_k \left\| k \right\|^r(2\nu + d_{\text{eff}}(l)) f_k f_k^* < +\infty,$$

where $\nu = 1/2 (3/2)$ for the NTK (RFK) and $k$ denotes the meta-patch $k_{i_l+1 \ldots, L+1}$ without the subscript to ease notation. Since the eigenvalues depend on the norm of $k$, Eq. 17 is equivalent to a finite-norm condition for all the derivatives of $f^*$ up to order $2m < r(2\nu + d_{\text{eff}}(l))$, \hspace{1cm} \hspace{1cm} (17)

\[ \|\Delta^{m/2} f^*\|^2 = \sum_k \| k \|^r(2\nu + d_{\text{eff}}(l)) f_k f_k^* < +\infty \] with $\Delta$ denoting the Laplace operator. As a result, if $f^*$ has derivatives of finite norm up to the $m$-th, then the source exponent can be tuned to $r = 2m/(2\nu + d_{\text{eff}}(l))$, inversely proportional to the effective dimensionality of $f^*$. Since the exponent on the right-hand side of Eq. 15 is an increasing function of $r$, the smaller the effective dimensionality of $f^*$ the faster the decay of the error—hence hierarchical kernels are adaptive to the spatial structure of $f^*$. For instance, a Lipschitz assumption on $f^*$ would lead to a fast rate $n^{-2/(2 + 2\nu + d_{\text{eff}}(l))}$ when $d = d_{\text{eff}}(L) \to \infty$. We expect the same result to hold for $s \geq 3$.

## 5 Asymptotics of generalisation in teacher-student scenarios

To further elucidate the interplay between the spatial structure of the architecture and that of data, we consider a stylised model of learning structured Gaussian random data with kernel regression \[35, 36, 39, 42, 43\]. Specifically, the target function is assumed to be a Gaussian random field $f^*(x)$ with covariance given by a teacher kernel $K_T$. Learning is performed with another kernel $K_S$—the student—via ridgeless regression. Notice that, in the ridgeless limit, the predictor of kernel regression coincides with the posterior mean obtained performing Bayesian inference under the prior that the target function is a Gaussian random field with covariance given by the student kernel \[38\]. Moreover, when $K_T$ and $K_S$ are chosen to be, respectively, the RFK and NTK of two neural networks, this scenario corresponds to an infinitely-wide neural network trained on the output of another infinitely-wide network with random weights. Here we consider deep hierarchical kernels, which allow us to generate target functions with a hierarchical structure: a single scenario is then specified by the depths and sets of filter sizes of the teacher and student kernels. After generating $f^*$ and a training set of size $n$, the predictor $f^n$ is given by Eq. 11 with $y_{\mu} = f^*(x_{\mu})$. The error $\epsilon(f^n)$ can then be averaged not only over training sets of fixed size, but also over different realisations of the target function $f^*$.

**Spectral bias ansatz.** In contrast with the optimally-regularised setting, where it is possible to apply the result discussed in the previous section, in the ridgeless case—where the correspondence between kernel methods and infinitely-wide neural networks actually holds—there are no rigorous results for the decay of the generalisation error. Therefore, we provide a heuristic derivation of the error decay based on a spectral bias ansatz. Consider the projections of the target function $f^*$ on the eigenfunctions of the student kernel $\tilde{Y}_K (f_k^\star)$ and assume that kernel methods learn only the $n$ projections corresponding to the highest eigenvalues. Then, if the decay of $f_k^\star$ with $k$ is sufficiently slow, one has

$$\tau(n) \sim \sum_{k \text{ s.t. } \Lambda_k^\star < \Lambda^{\text{eff}}(n)} \mathbb{E}_{f^*} \left[ (f_k^\star)^2 \right],$$

\hspace{1cm} \hspace{1cm} (18)

\[ We are again limiting the presentation to the case $s = 2$ but the extension to the general case is immediate.\]
with \( A^S(n) \) the value of the \( n \)-th largest eigenvalue of the student. This result can be derived using the replica method of statistical physics \([23][39][24]\) (details in App. F) or by assuming that input points lie on a lattice \([3]\). We validate all predictions based on Eq. 18 with numerical experiments.

**Shallow architectures.** This method was already used in \([14]\) in the shallow case \( L = 1 \) with a different distribution of input points. In our setting, with both teacher and student having filter size \( s \),

\[
\tau(n) \sim n^{\frac{2\nu_T}{\nu_T - 1}}, \quad \nu_T = 1/2 (3/2) \text{ for an NTK (RFK) teacher.} \tag{19}
\]

One can compare this result with the rigorous bound in the optimally-regularised case \([\text{eq. } 13]\). In particular, the capacity condition is satisfied with \( \alpha = 1 + 2\nu_s/s \), and the source condition is satisfied setting \( r = 2\nu_T/(2\nu_T + s) \). Therefore, \( \tau(n) \leq C'n^{-\beta} \), with \( \beta = 2\nu_T/(2\nu_T + s - 1) \) (see App. F for details). Since these exponents are independent of the ambient dimension \( d \), the curse of dimensionality is beaten. This success comes as a price: the network is fine-tuned to the task and is only able to approximate local functions that do not couple different patches of the input.

**Deep architectures.** This constraint can be lifted thanks to depth: consider a student of depth \( L + 1 \) with filter sizes \( (s_1, \ldots, s_L) \) and \( p_L = 1 \), such that \( d = \prod_{1 \leq l \leq L} s_l \). Let the teacher be a hierarchical kernel of depth \( l + 1 \), \( l \leq L \), with filter sizes equal to those of the student up to \( s_l \). The scaling of the \( n \)-th largest student eigenvalue with \( n \) is always determined by the largest sector, but the only non-zero projections of \( f^* \) stem from the sectors with \( l' \leq l \). Plugging in Eq. 18 the relevant decays from Thm. 3.1 we obtain (details in App. F)

\[
\tau(n) \sim n^{\frac{2\nu_T - 2\nu_s + d_{\text{eff}}(L)}{\nu_T - d_{\text{eff}}(L)}}, \tag{20}
\]

with \( \nu_S = 1/2 (3/2) \) when the student kernel is an NTK (RFK). Notice that, in taking the limit \( d \to \infty \) while keeping the filter sizes up to \( s_l \) finite, the exponent of the error decay does not go to zero—as if it was cursed by the dimensionality—but converges to \( 2\nu_T/(2\nu_S + d_{\text{eff}}(l)) \). This is only slightly worse than \( 2\nu_T/d_{\text{eff}}(l) \) obtained when the student is matched to the teacher (compare with Eq. 19) with \( s = 1 \) replaced by \( d_{\text{eff}}(l) \). In comparison, the source-capacity bound results in

\[
\beta = \frac{2\nu_T(2\nu_S + d_{\text{eff}}(L))}{2\nu_T(2\nu_S + d_{\text{eff}}(L)) + (2\nu_S + d_{\text{eff}}(l)d_{\text{eff}}(L))} \xrightarrow{d \to \infty} \frac{2\nu_T}{2\nu_S + 2\nu_T + d_{\text{eff}}(l)}, \tag{21}
\]

which also does not vanish with the full dimensionality. These results show that hierarchical kernels play significantly better with the approximation-estimation trade-off, as they are able to approximate global functions of the input while not being cursed when the target function has a local structure.

**Curse of dimensionality for hierarchical targets** By contrast, if the teacher is as deep as the student and global, i.e. \( p_L = 1 \), the resulting rate is \( n^{-2\nu_T/d_{\text{eff}}(L)} \), ‘cursed’ by the effective dimensionality \( d_{\text{eff}}(L) = p \times (s - 1) \) which diverges together with \( d = p \times s \) in the \( d \to \infty \) limit. Hence having a global target function with a hierarchical structure is not sufficient to beat the curse of dimensionality, as it leads to a similar rate to the one obtained with a non-hierarchical target. In fact, because of the equivalence of the predictors of kernel ridgeless regression and Bayesian inference, the rate obtained in the case where teacher and student kernel match is Bayes-optimal. We conclude that it is impossible to beat the curse of dimensionality with any method when learning the output of an infinitely-wide, deep hierarchical network with random weights. Therefore, target functions of this kind cannot be a good model of learnable tasks.

**Numerical experiments.** Fig. 2 confirms the picture emerging from our calculations. In particular, panel (a) shows a depth-four student learning depth-two, depth-three, and depth-four teachers. This student is not cursed in the first two cases and is cursed in the third one, which corresponds to a global target function. Panel (b) illustrates the curse of dimensionality with the effective input dimension \( d_{\text{eff}} \) by comparing the learning curves of depth-three students learning global target functions with an increasing number of variables. All our theoretical predictions are in excellent agreement with the simulations. All the details of numerical experiments are reported in App. G together with additional results for \( s_1 \geq 3 \) (Fig. S2), a comparison between the ridgeless and optimally-regularised cases (Fig. S3), and experiments with the CIFAR-10 dataset (Fig. S4).
6 Conclusions and outlook

We have proved that deep CNNs can adapt to the spatial scale of the target function, thus beating the curse of dimensionality if the target depends only on local groups of variables. Yet, if considered as ‘teachers’, they generate functions that cannot be learnt efficiently in high dimensions, even in the Bayes-optimal setting where the student is matched to the teacher. Thus, the architectures we considered are not good models of the hierarchical structure of real data which are efficiently learnable.

Enforcing a stronger notion of compositionality is an interesting endeavour for the future. Following [8], one may consider a much smaller family of functions of the form

\[ f_{i_{l+1}}(x_{i_{l+1}}) = f_{i_{l+1}}(f_{i_{l+1}}(\ldots f_{i_{l+1}}(x_{i_1}))) \]  

\[ (22) \]

Preliminary results assuming that the functions \( f \)'s are random Gaussian fields suggest that such a task is not learnable efficiently by a hierarchical CNN in the kernel regime. It is unclear whether this remains true when the network is trained in a regime where features can be learnt from data [40, 41]. Recent works, for instance, have established that under certain conditions locality can be learnt from scratch [42]. Can compositionality be learnt too?

Finally, another additional direction to explore is the stability of the task toward smooth transformations or diffeomorphisms. This form of stability has been proposed as a key element to understanding how the curse of dimensionality is beaten for image datasets [43, 44]. Such a property can be enforced with pooling operations [21, 45]; therefore diagonalising the NTK in this case as well would be of high interest.

Acknowledgements

We thank Massimo Sorella for pointing to us the relationship between differentiability and asymptotics of the Fourier coefficients. We also thank Antonio Sclocchi, Leonardo Petrini, Umberto Maria Tomasini and Marcello d’Abbicco for helpful discussions. This work was supported by a grant from the Simons Foundation (#454953 Matthieu Wyart).

References

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

[2] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. arXiv preprint arXiv:1712.00409, 2017.
[3] Stefano Spigler, Mario Geiger, and Matthieu Wyart. Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124001, 2020.

[4] Lechao Xiao and Jeffrey Pennington. Synergy and symmetry in deep learning: Interactions between the data, model, and inference algorithm. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 24347–24369. PMLR, 17–23 Jul 2022.

[5] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987.

[6] Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning*. PMLR, 2018.

[7] Hao Henry Zhou, Yunyang Xiong, and Vikas Singh. Building bayesian neural networks with blocks: On structure, interpretability and uncertainty. *arXiv preprint arXiv:1806.03563*, 2018.

[8] Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

[9] Arturo Deza, Qianli Liao, Andrzej Banburski, and Tomaso Poggio. Hierarchically compositional tasks and deep convolutional networks. *arXiv preprint arXiv:2006.13915*, 2020.

[10] Michael Kohler, A Krzyzak, and Benjamin Walter. On the rate of convergence of image classifiers based on convolutional neural networks. *arXiv preprint arXiv:2003.01526*, 2020.

[11] Tomaso Poggio, Andrzej Banburski, and Qianli Liao. Theoretical issues in deep networks. *Proceedings of the National Academy of Sciences*, 117(48):30039–30045, 2020.

[12] Eran Malach and Shai Shalev-Shwartz. Computational separation between convolutional and fully-connected networks. In *International Conference on Learning Representations*, 2021.

[13] Alberto Bietti. Approximation and learning with deep convolutional models: a kernel perspective. In *International Conference on Learning Representations*, 2022.

[14] Alessandro Favero, Francesco Cagnetta, and Matthieu Wyart. Locality defeats the curse of dimensionality in convolutional teacher-student scenarios. In *Advances in Neural Information Processing Systems*, 34, 2021.

[15] Theodor Misiakiewicz and Song Mei. Learning with convolution and pooling operations in kernel methods. *arXiv preprint arXiv:2111.08308*, 2021.

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

[17] Amnon Geifman, Meirav Galun, David Jacobs, and Ronen Basri. On the spectral bias of convolutional neural tangent and gaussian process kernels. *arXiv preprint arXiv:2203.09255*, 2022.

[18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[19] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

[20] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

[21] Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. *Advances in Neural Information Processing Systems*, 32, 2019.

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

[23] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR, 2020.
[24] Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151, 2021.

[25] Alberto Bietti and Francis Bach. Deep equals shallow for relu networks in kernel regimes. In *ICLR 2021–International Conference on Learning Representations*, pages 1–22, 2021.

[26] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

[27] Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. *Advances In Neural Information Processing Systems*, 29, 2016.

[28] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *Advances in Neural Information Processing Systems*, 32, 2019.

[29] James Mercer. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209(441-458):415–446, 1909.

[30] Harold Widom. Asymptotic behavior of the eigenvalues of certain integral equations. *Transactions of the American Mathematical Society*, 109(2):278–295, 1963.

[31] Douglas Azevedo and Valdir A Menegatto. Eigenvalues of dot-product kernels on the sphere. *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics*, 3(1), 2015.

[32] Meyer Scetbon and Zaid Harchaoui. A spectral analysis of dot-product kernels. In *International Conference on Artificial Intelligence and Statistics*, pages 3394–3402. PMLR, 2021.

[33] Francis Bach. Learning theory from first principles, 2021.

[34] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[35] Peter Sollich. Learning curves for gaussian processes. *Advances in neural information processing systems*, 11, 1998.

[36] Peter Sollich. Gaussian process regression with mismatched models. *Advances in Neural Information Processing Systems*, 14, 2001.

[37] Jonas Paccolat, Stefano Spigler, and Matthieu Wyart. How isotropic kernels perform on simple invariants. *Machine Learning: Science and Technology*, 2(2):025020, 2021.

[38] George S Kimeldorf and Grace Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502, 1970.

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

[40] Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3655–3673. PMLR, 2020.

[41] Mario Geiger, Leonardo Petrini, and Matthieu Wyart. Landscape and training regimes in deep learning. *Physics Reports*, 924:1–18, 2021.

[42] Alessandro Ingrosso and Sebastian Goldt. Data-driven emergence of convolutional structure in neural networks. *arXiv preprint arXiv:2202.00565*, 2022.

[43] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

[44] Leonardo Petrini, Alessandro Favero, Mario Geiger, and Matthieu Wyart. Relative stability toward diffeomorphisms indicates performance in deep nets. *Advances in Neural Information Processing Systems*, 34, 2021.
[45] Alberto Bietti, Luca Venturi, and Joan Bruna. On the sample complexity of learning under geometric stability. *Advances in Neural Information Processing Systems*, 34:18673–18684, 2021.

[46] Alex Smola, Zoltán Ovári, and Robert C Williamson. Regularization with dot-product kernels. *Advances in neural information processing systems*, 13, 2000.

[47] Costas Efthimiou and Christopher Frye. *Spherical harmonics in p dimensions*. World Scientific, 2014.

[48] Kendall Atkinson and Weimin Han. *Spherical harmonics and approximations on the unit sphere: an introduction*, volume 2044. Springer Science & Business Media, 2012.

[49] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

[50] Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems 22*, pages 342–350. Curran Associates, Inc., 2009.

[51] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

[52] Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Implicit regularization of random feature models. In *International Conference on Machine Learning*, pages 4631–4640. PMLR, 2020.

[53] Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Kernel alignment risk estimator: Risk prediction from training data. *Advances in Neural Information Processing Systems*, 33, 2020.

[54] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.
Supplementary material

A Harmonic analysis on the sphere

A.1 Dot-product kernels on the sphere ............................................. 15
A.2 Multi-dot-product kernels on the multi-sphere ............................ 16

B RFK and NTK of deep convolutional networks ............................ 17

C Spectra of deep convolutional kernels ....................................... 17
C.1 Singular expansion of hierarchical kernels ................................. 18
C.2 Patches on the ring ............................................................... 20
C.3 Patches on the s-dimensional hypersphere ............................... 21
C.4 General depth ................................................................. 24

D Generalisation bounds for kernel regression and spatial adaptivity 24

E Statistical mechanics of generalisation in kernel regression ........... 25

F Learning curves of teacher-student scenarios ............................... 26
F.1 Recap of error rates ............................................................. 26
F.2 Shallow architectures .......................................................... 28
F.3 Deep architectures ............................................................. 29
F.4 Source-capacity conditions ................................................... 29

G Numerical experiments .......................................................... 30
G.1 Experimental setup ............................................................ 30
G.2 Teacher-student learning curves ............................................. 30
G.3 CIFAR-2 learning curves ...................................................... 31

A Harmonic analysis on the sphere

This appendix collects some introductory background on spherical harmonics and dot-product kernels on the sphere \( [46] \). See \([47,48,49]\) for a complete description. Spherical harmonics are homogeneous polynomials on the sphere \( S^{s-1} = \{ x \in \mathbb{R}^s \mid \| x \| = 1 \} \), with \( \| \cdot \| \) denoting the L2 norm. Given the polynomial degree \( k \in \mathbb{N} \), there are \( N_{k,s} \) linearly independent spherical harmonics of degree \( k \) on \( S^{s-1} \), with

\[
N_{k,s} = \frac{2k + s - 2}{k} \binom{s + k - 3}{k - 1},
\]

\( \{ N_{0,d} = 1 \quad \forall d, \}
\]

\[
N_{k,d} \sim k^{d-2} \quad \text{for } k \gg 1.
\]

Thus, we can introduce a set of \( N_{k,s} \) spherical harmonics \( Y_{k,\ell} \) for each \( k \), with \( \ell \) ranging in \( 1, \ldots, N_{k,s} \), which are orthonormal with respect to the uniform measure on the sphere \( d\tau(x) \),

\[
\langle Y_{k,\ell}, Y_{k,\ell'} \rangle_{S^{s-1}} := \int_{S^{s-1}} d\tau(x) Y_{k,\ell}(x) Y_{k,\ell'}(x) = \delta_{\ell,\ell'}.
\]

Because of the orthogonality of homogeneous polynomials with different degree, the set \( \{ Y_{k,\ell} \}_{k,\ell} \) is a complete orthonormal basis for the space of square-integrable functions on the \( s \)-dimensional unit sphere. Furthermore, spherical harmonics are eigenfunctions of the Laplace-Beltrami operator \( \Delta \), which is nothing but the restriction of the standard Laplace operator to \( S^{s-1} \).

\[
\Delta Y_{k,\ell} = -k(k+s-2)Y_{k,\ell}.
\]

(S1)

(S2)

(S3)
The Laplace-Beltrami operator $\Delta$ can also be used to characterise the differentiability of functions $f$ on the sphere via the $L^2$ norm of some power of $\Delta$ applied to $f$.

By fixing a direction $y$ in $\mathbb{S}^{d-1}$ one can select, for each $k$, the only spherical harmonic of degree $k$ which is invariant for rotations that leave $y$ unchanged. This particular spherical harmonic is, in fact, a function of $x \cdot y$ and is called the Legendre polynomial of degree $k$, $P_{k,s}(x \cdot y)$ (also referred to as Gegenbauer polynomial). Legendre polynomials can be written as a combination of the orthonormal spherical harmonics $Y_{k,\ell}$ via the addition formula [48, Thm. 2.9],

$$P_{k,s}(x \cdot y) = \frac{1}{N_{k,s}} \sum_{\ell=1}^{N_{k,s}} Y_{k,\ell}(x)Y_{k,\ell}(y). \quad (S4)$$

Alternatively, $P_{k,s}$ is given explicitly as a function of $t = x \cdot y \in [-1, +1]$ via the Rodrigues’ formula [48, Thm. 2.23],

$$P_{k,s}(t) = (-1)^k \left( \frac{1}{2} \right)^k \frac{\Gamma \left( \frac{s+1}{2} \right)}{\Gamma \left( \frac{k+s+1}{2} \right)} \left( 1 - t^2 \right)^{\frac{s-k}{2}} \frac{d^k}{dt^k} \left( 1 - t^2 \right)^{k+s-1}. \quad (S5)$$

Legendre polynomials are orthogonal on $[-1, +1]$ with respect to the measure with density $(1 - t^2)^{(s-3)/2}$, which is the probability density function of the scalar product between to points on $\mathbb{S}^{s-1}$.

$$\int_{-1}^{+1} dt \left( 1 - t^2 \right)^{\frac{s-3}{2}} P_{k,s}(t)P_{k',s}(t) = \frac{\left( s-1 \right)}{\left( s-2 \right)} \frac{\delta_{k,k'}}{N_{k,s}}, \quad (S6)$$

with $\mathbb{S}^{s-1}$ denoting the surface area of the $s$-dimensional unit sphere.

To sum up, given $x, y \in \mathbb{S}^{s-1}$, functions of $x$ or $y$ can be expressed as a sum of projections on the orthonormal spherical harmonics $\{Y_{k,\ell}\}_{k,\ell}$, whereas functions of $x \cdot y$ can be expressed as a sum of projections on the Legendre polynomials $\{P_{k,s}(x \cdot y)\}_{k,s}$. The relationship between the two expansions is elucidated in the Funk-Hecke formula [48, Thm. 2.22],

$$\int_{\mathbb{S}^{s-1}} d\tau(y) f(x \cdot y)Y_{k,\ell}(y) = Y_{k,\ell}(x) \frac{\left( s-2 \right)}{\left( s-1 \right)} \int_{-1}^{+1} dt \left( 1 - t^2 \right)^{\frac{s-3}{2}} f(t)P_{k,s}(t). \quad (S7)$$

If the function $f$ has continuous derivatives up to the $k$-th order in $[-1, +1]$, then one can plug Rodrigues’ formula in the right-hand side of Funk-Hecke formula and get, after $k$ integrations by parts,

$$\int_{\mathbb{S}^{s-1}} d\tau(y) f(x \cdot y)Y_{k,\ell}(y) = Y_{k,\ell}(x) \frac{\left( s-2 \right)}{\left( s-1 \right)} \frac{\Gamma \left( \frac{s+1}{2} \right)}{2^k \Gamma \left( k+\frac{s+1}{2} \right)} \int_{-1}^{+1} dt f^{(k)}(t) \left( 1 - t^2 \right)^{k+s-1}, \quad (S8)$$

with $f^{(k)}(t)$ denoting the $k$-th order derivative of $f$ in $t$. This trick also applies to functions which are not $k$ times differentiable at $\pm 1$, provided the boundary terms due to integration by parts vanish.

### 1. Dot-product kernels on the sphere

Dot-product kernels are kernels which depend on the two inputs $x$ and $y$ via their scalar product $x \cdot y$. When the inputs lie on the unit sphere $\mathbb{S}^{s-1}$, one can use the machinery introduced in the previous section to arrive immediately at the Mercer’s decomposition of the kernel [46].

$$K(x \cdot y) = \sum_{k \geq 0} \left( \frac{\left( s-2 \right)}{\left( s-1 \right)} \frac{\Gamma \left( \frac{s+1}{2} \right)}{2^k \Gamma \left( k+\frac{s+1}{2} \right)} \int_{-1}^{+1} dt \left( 1 - t^2 \right)^{\frac{s-3}{2}} K(t)P_{k,s}(t) \right) P_{k,s}(x \cdot y)$$

$$= \sum_{k \geq 0} \left( \frac{\left( s-2 \right)}{\left( s-1 \right)} \frac{\Gamma \left( \frac{s+1}{2} \right)}{2^k \Gamma \left( k+\frac{s+1}{2} \right)} \int_{-1}^{+1} dt \left( 1 - t^2 \right)^{\frac{s-3}{2}} K(t)P_{k,s}(t) \right) \sum_{\ell=1}^{N_{k,s}} Y_{k,\ell}(x)Y_{k,\ell}(y) \quad (S9)$$

In the first line we have just decomposed $K$ into projections onto the Legendre polynomials, the second line follows immediately from the addition formula, the third is just a definition of the
eigenvalues $\Lambda_k$. Notice that the eigenfunctions of the kernel are orthonormal spherical harmonics and the eigenvalues are degenerate with respect to the index $\ell$. The Reproducing Kernel Hilbert Space (RKHS) of $\mathcal{K}$ can be characterised as follows,

$$\mathcal{H} = \left\{ f : \mathbb{S}^{s-1} \to \mathbb{R} \text{ s. t. } \|f\|_{\mathcal{H}} := \sum_{k \geq 0, \Lambda_k \neq 0} \sum_{\ell = 1}^{N_{k,s}} \frac{\langle f, Y_{k,\ell} \rangle_{\mathbb{S}^{s-1}}}{\Lambda_k} < +\infty \right\}. \quad (S10)$$

### A.2 Multi-dot-product kernels on the multi-sphere

The Mercer’s decomposition of dot-product kernels extends naturally to the case considered in this paper, where the input space is the Cartesian product of $p$ $s$-dimensional unit sphere,

$$\mathcal{X}_{s,p} = \{ \mathbf{x} = (x_1, \ldots, x_p) \mid x_i \in \mathbb{S}^{s-1} \forall i = 1, \ldots, p \} = \prod_{i=1}^{p} \mathbb{S}^{s-1} \quad (S11)$$

which we refer to as the **multi-sphere** following the notation of [17]. After defining a scalar product between functions on $\mathcal{X}_{s,p}$ by direct extension of Eq. (S2) one can immediately find a set of orthonormal polynomials by taking products of spherical harmonics. With the multi-index notation $k = (k_1, \ldots, k_p)$, $\ell = (\ell_1, \ldots, \ell_p)$, for all $\mathbf{x} \in \mathcal{X}_{s,p}$

$$\tilde{Y}_{k,\ell}(\mathbf{x}) = \prod_{i=1}^{p} Y_{k_i,\ell_i}(x_i), \text{ with } k_i \geq 0 \text{ and } \ell_i = 1, \ldots, N_{k_i,s} = \frac{2k_i + s - 2}{k_i} \left( s + k_i - 3 \right). \quad (S12)$$

These product spherical harmonics $\tilde{Y}_{k,\ell}(\mathbf{x})$ span the space of square-integrable functions on $\mathcal{X}_{s,p}$. Furthermore, as each spherical harmonic is an eigenfunction of the Laplace-Beltrami operator, $\tilde{Y}_{k,\ell}$ is an eigenfunction of the sum of Laplace-Beltrami operators on the $p$ unit spheres,

$$\Delta_{p,s} \tilde{Y}_{k,\ell} := \left( \sum_{i=1}^{p} \Delta_i \right) \prod_{i=1}^{p} Y_{k_i,\ell_i} = \left( \sum_{i=1}^{p} (-\ell_i)(\ell_i + s - 2) \right) \tilde{Y}_{k,\ell}. \quad (S13)$$

We can thus characterise the differentiability of functions of the multi-sphere $\mathcal{X}_{s,p}$ via finiteness in $L^2$ norm of some power of $\Delta_{p,s}$.

Similarly, we can consider products of Legendre polynomials to obtain a set of orthogonal polynomials on $[-1, 1]^p$ (see [17], appendix A). Then, any function $f$ on $\mathcal{X}_{s,d} \times \mathcal{X}_{s,d}$ which depends only on the $p$ scalar products between patches,

$$f(\mathbf{x}, \mathbf{y}) = g(x_1 \cdot y_1, \ldots, x_p \cdot y_p), \quad (S14)$$

can be written as a sum of projections on products of Legendre polynomials

$$\tilde{P}_{k,s}(t) := \prod_{i=1}^{p} P_{k_i,s}(t_i). \quad (S15)$$

Following again [17], we call such functions **multi-dot-product** kernels. When fixing one of the two arguments of $f$ (say $\mathbf{x}$), $f$ becomes a function on $\mathcal{X}_{s,d} \times \mathcal{X}_{s,d}$ and can be written as a sum of projections on the $\tilde{Y}_{k,\ell}$’s. The two expansions are related by the following generalised Funk-Hecke formula,

$$\left( \prod_{i=1}^{p} \int_{\mathbb{S}^{s-1}} d\tau(y_i) \right) g(x_1 \cdot y_1, \ldots, x_p \cdot y_p) \tilde{Y}_{k,\ell}(y) = \tilde{Y}_{k,\ell}(y) \left( \frac{|\mathbb{S}^{s-2}|}{|\mathbb{S}^{s-1}|} \right)^p \left( \prod_{i=1}^{p} \int_{-1}^{+1} dt_i \left( 1 - t_i^2 \right)^{\frac{p-2}{2}} \tilde{P}_{k_i,s}(t_i) \right) g(t_1, \ldots, t_p). \quad (S16)$$

Having introduced the product spherical harmonics $\tilde{Y}_{k,\ell}$ as basis of $\mathcal{X}_{s,p}$ and the product Legendre polynomials $\tilde{P}_{k,s}(t)$ as basis of $[-1, +1]^p$, the Mercer’s decomposition of multi-dot-product kernels
follows immediately.

\[
\mathcal{K} (\{x_i \cdot y_i\}_i) = \sum_{k \geq 0} \left( \prod_{i=1}^{p} N_{h_i, s} \right) \int_{-1}^{1} \frac{dt_i (1 - t_i^2)^{s-1}}{\pi} P_k, s (t_i) \mathcal{K} (\{t_i\}_i) P_k, s (\{x_i \cdot y_i\}_i) \\
= \sum_{k \geq 0} \sum_{\ell=1}^{N_k} Y_{k, \ell}(x) Y_{k, \ell}(y).
\]

(S17)

**B RFK and NTK of deep convolutional networks**

This appendix gives an extended version of Def. 3.1 including the full form of the NTK of hierarchical CNNs. We refer the reader to [28] for the derivation.

**Definition B.1 (RFK and NTK of hierarchical CNNs)** Let \(x, y \in \mathcal{X} = \prod_{i=1}^{p} S^{s_i - 1}\). Denote tuples of the kind \(i_1 i_2 \ldots i_m\) with \(i_{m+1}\) for \(m \geq 1\). For \(m < l\), \(i_{l-m}\) denotes the empty tuple. For each tuple \(i_{2 \rightarrow L+1}\), denote with \(t_{i_{2 \rightarrow L+1}}\) the scalar product between the \(s\)-dimensional patches of \(x\) and \(y\) identified by the same tuple, i.e.

\[
t_{i_{2 \rightarrow L+1}} = \langle x_{i_{2 \rightarrow L+1}}, y_{i_{2 \rightarrow L+1}} \rangle
\]

For \(1 \leq l \leq L + 1\), denote with \(\{t_{i_{2 \rightarrow L+1}}\}_i\) the sequence of \(t\)'s obtained by letting the indices of the tuple \(i_{2 \rightarrow L+1}\) vary in their respective range. Consider a hierarchical CNN with \(L\) hidden layers, filter sizes \((s_1, \ldots, s_L)\), \(p_l \geq 1\) and all the weights \(w^{(l)}_{h,i}, w^{(l)}_{h,h'}, w^{(L+1)}_{h,i}\) initialised as Gaussian random numbers with zero mean and unit variance.

**RFK**. The corresponding RFK (or covariance kernel) is a function \(K^{(L+1)}_{RFK}\) of the \(p_1 = d / s_1\) scalar products \(t_{i_{2 \rightarrow L+1}}\) which can be obtained recursively as follows. With \(\kappa_1(t) = ((\pi - \arccos t) t + \sqrt{1 - t^2}) / \pi\),

\[
K^{(1)}_{RFK} (t_{i_{2 \rightarrow L+1}}) = \kappa_1 (t_{i_{2 \rightarrow L+1}});
\]

\[
K^{(l)}_{RFK} (\{t_{i_{2 \rightarrow L+1}}\}_i) = \kappa_1 \left( \frac{1}{s_l} \sum_{i_{l-1}} K^{(l-1)}_{RFK} (\{t_{i_{2 \rightarrow L+1}}\}_{i_{2 \rightarrow L+1}}) \right); \\ \forall l \in [2 \ldots L] \text{ if } L > 1;
\]

\[
K^{(L+1)}_{RFK} (\{t_{i_{2 \rightarrow L+1}}\}_i) = \frac{1}{p_L} \sum_{i_{L+1}=1}^{p_L} K^{(L)} (\{t_{i_{2 \rightarrow L+1}}\}_i).
\]

(S19)

**NTK**. The NTK of the same hierarchical CNN is also a function of the \(p_1 = d / s_1\) scalar products \(t_{i_{2 \rightarrow L+1}}\) which can be obtained recursively as follows. With \(\kappa_0(t) = (\pi - \arccos t) / \pi\),

\[
\mathcal{N} \mathcal{T} \mathcal{K}^{(1)} (t_{i_{2 \rightarrow L+1}}) = \kappa_1 (t_{i_{2 \rightarrow L+1}}) + (t_{i_{2 \rightarrow L+1}}) \kappa_0 (t_{i_{2 \rightarrow L+1}});
\]

\[
\mathcal{N} \mathcal{T} \mathcal{K}^{(l)} (\{t_{i_{2 \rightarrow L+1}}\}_i) = \mathcal{R} \mathcal{F} \mathcal{K}^{(l-1)} (\{t_{i_{2 \rightarrow L+1}}\}_{i_{2 \rightarrow L+1}}) + \left( \frac{1}{s_l} \sum_{i_{l-1}} \mathcal{N} \mathcal{T} \mathcal{K}^{(l-1)} (\{t_{i_{2 \rightarrow L+1}}\}_{i_{2 \rightarrow L+1}}) \right) \\
\times \kappa_0 \left( \frac{1}{s_l} \sum_{i_{l-1}} \mathcal{R} \mathcal{F} \mathcal{K}^{(l-1)} (\{t_{i_{2 \rightarrow L+1}}\}_{i_{2 \rightarrow L+1}}) \right); \\ \forall l \in [2 \ldots L] \text{ if } L > 1;
\]

\[
\mathcal{N} \mathcal{T} \mathcal{K}^{(L+1)} (\{t_{i_{2 \rightarrow L+1}}\}_i) = \frac{1}{p_L} \sum_{i_{L+1}=1}^{p_L} \mathcal{N} \mathcal{T} \mathcal{K}^{(L)} (\{t_{i_{2 \rightarrow L+1}}\}_i).
\]

(S20)

**C Spectra of deep convolutional kernels**

In this section we provide a proof to Thm. 3.1. Our strategy is to relate the asymptotic decay of eigenvalues to the singular behaviour of the kernel, as it is customary in Fourier analysis and was
As a result, hierarchical kernels have a singular expansion when the

\[ \lim_{t \to 0} \kappa(t) = \begin{cases} (\pi - \arccos t) / \pi, & \kappa(t) = (\pi - \arccos t) t + \sqrt{1 - t^2} / \pi. \end{cases} \]  

(S21)

The functions \( \kappa_0 \) and \( \kappa_1 \) are non-analytic in \( t = \pm 1 \), with the following singular expansion [25].

Near \( t = 1 \), with \( u = 1 - t \)

\[ \begin{cases} \kappa_0(1 - u) = 1 - \frac{\sqrt{2}}{\pi} u^{1/2} + O(u^{3/2}), \\ \kappa_1(1 - u) = 1 - u + \frac{2\sqrt{2}}{3\pi} u^{3/2} + O(u^{5/2}). \end{cases} \]  

(S22)

Near \( t = -1 \), with \( u = 1 + t \),

\[ \begin{cases} \kappa_0(-1 + u) = \frac{\sqrt{2}}{\pi} u^{1/2} + O(u^{3/2}), \\ \kappa_1(-1 + u) = \frac{2\sqrt{2}}{3\pi} u^{3/2} + O(u^{5/2}). \end{cases} \]  

(S23)

As a result, hierarchical kernels have a singular expansion when the \( t_{i_2 \to L+1} \)'s are close to \( \pm 1 \). In particular, the following expansions are relevant for computing the asymptotic scaling of eigenvalues.

**Proposition C.1 (RFK when \( x = y \))** The RFK of a hierarchical network of depth \( L + 1 \), filter sizes \( \{s_1, \ldots, s_L\} \) and \( p_L \geq 1 \) has the following singular expansion when all \( t_{i_2 \to L+1} \to 1 \). With \( u_{i_2 \to L+1} = 1 - t_{i_2 \to L+1}, c = 2\sqrt{2} / (3\pi), \) and \( \prod_{l \in I} s_{l+1} = 1 \) if \( I \) is the empty set,

\[ \kappa_{\text{RFK}}^{(L+1)} \left( \left\{ 1 - u_{i_2 \to L+1} \right\}_{i_2 \to L+1} \right) = 1 - \frac{1}{p_{L}} \sum_{i_2 \to L+1} u_{i_2 \to L+1} \]

\[ + \frac{c}{p_{L}} \sum_{l' = 1}^L \left( \prod_{l' + 1 \leq L+1} s_{l'} \right) \sum_{i_2 \to L+1} \left( \prod_{2 \leq l'' < L} s_{l''} \right) \sum_{i_2 \to L+1} u_{i_2 \to L+1} \]

\[ + O(u_{i_2 \to L+1}^{5/2}). \]  

(S24)

**Proof.** With \( L = 1 \) one has, from Eq. 8 (recall that \( i_{2 \to 1+1} = i_{2 \to 2} \) reduces to a single index)

\[ \kappa_{\text{RFK}}^{(1)} (1 - u_{i_2}) = 1 - u_{i_2} + c u_{i_2}^{3/2} + O(u_{i_2}^{5/2}) \Rightarrow \]

\[ \kappa_{\text{RFK}}^{(1+1)} \left( \left\{ 1 - u_{i_2} \right\}_{i_2} \right) = 1 - \frac{1}{p_{1}} \sum_{i_2} u_{i_2} + \frac{c}{p_{1}} \sum_{i_2} u_{i_2}^{3/2} + O(u_{i_2}^{5/2}). \]  

(S25)
Thus, for \( \kappa \) so that

\[
K_u(L^{(2)}) = \kappa_1 \left( \frac{1}{s_2} \sum_{i_2} u_{i_2,i_3} + \frac{c}{s_2} \sum_{i_2} u_{i_2,i_3}^{3/2} + O(u_{i_2,i_3}^{5/2}) \right)
\]

\[
= 1 - \frac{1}{s_2} \sum_{i_2} u_{i_2,i_3} + \frac{c}{s_2} \sum_{i_2} u_{i_2,i_3}^{3/2} + c \left( \frac{1}{s_2} \sum_{i_2} u_{i_2,i_3} \right)^{3/2} + O(u_{i_2,i_3}^{5/2}),
\]

therefore

\[
K_{\text{RFK}}^{(2+1)} \left( \{1 - u_{i_2,i_3}\}_{i_2,i_3} \right) = 1 - \frac{1}{s_2p_2} \sum_{i_2,i_3} u_{i_2,i_3} + \frac{1}{p_2} \sum_{i_2,i_3} u_{i_2,i_3}^{3/2} + c \sum_{i_3} \left( \frac{1}{s_2} \sum_{i_2} u_{i_2,i_3} \right)^{3/2} + O(u_{i_2,i_3}^{5/2}).
\]

The proof of the general case follows by induction by applying the function \( \kappa_1 \) to the singular expansion of the kernel with \( L - 1 \) hidden layers, then using Eq. S22.

**Proposition C.2 (RFK when \( x = -y \))** The RFK of a hierarchical network of depth \( L + 1 \), filter sizes \( (s_1, \ldots, s_L) \) and \( p_L \geq 1 \) has the following singular expansion when all \( t_{i_2 \rightarrow L+1} \rightarrow -1 \). With \( u_{i_2 \rightarrow L+1} = 1 + t_{i_2 \rightarrow L+1}, c = 2\sqrt{2}/(3\pi) \) and \( \prod_{i \in I} s_i := 1 \) if \( I \) is the empty set,

\[
K_{\text{RFK}}^{(L+1)} \left( \{1 + u_{i_2 \rightarrow L+1}\}_{i_2 \rightarrow L+1} \right) = b_L + \left( \prod_{2 \leq i' \leq L} s_{i'} \right) \sum_{i_2 \rightarrow L+1} u_{i_2,i_3}^{3/2} + O(u_{i_2,i_3}^{5/2}),
\]

with \( b_L = \kappa_1(b_{L-1}), b_1 = 0; \) and \( c_L = c_{L-1} \kappa_1'(b_{L-1}), c_1 = c. \)

**Proof.** This can be proved again by induction. For \( L = 1 \),

\[
K_{\text{RFK}}^{(1)}(-1 + u_{i_2}) = cu_{i_2}^{3/2} + O(u_{i_2}^{5/2}) \Rightarrow K_{\text{RFK}}^{(1+1)} \left( \{1 + u_{i_2}\}_{i_2} \right) = \frac{c}{p_1} \sum_{i_2} u_{i_2}^{3/2} + O(u_{i_2}^{5/2}).
\]

Thus, for \( L = 2 \),

\[
K_{\text{RFK}}^{(2)} \left( \{1 + u_{i_2,i_3}\}_{i_2} \right) = \kappa_1 \left( \frac{c}{s_2} \sum_{i_2} u_{i_2,i_3}^{3/2} + O(u_{i_2,i_3}^{5/2}) \right) = \kappa_1(0) + \kappa_1'(0) \left( \frac{c}{s_2} \sum_{i_2} u_{i_2,i_3}^{3/2} \right) + O(u_{i_2,i_3}^{5/2}),
\]

so that

\[
K_{\text{RFK}}^{(2+1)} \left( \{1 + u_{i_2,i_3}\}_{i_2,i_3} \right) = \kappa_1(0) + \kappa_1'(0) \sum_{i_2,i_3} u_{i_2,i_3}^{3/2} + O(u_{i_2,i_3}^{5/2}).
\]

The proof is completed by applying the function \( \kappa_1 \) to the singular expansion of the kernel with \( L - 1 \) hidden layers.

**Proposition C.3 (NTK when \( x = y \))** The NTK of a hierarchical network of depth \( L + 1 \), filter sizes \( (s_1, \ldots, s_L) \) and \( p_L \geq 1 \) has the following singular expansion when all \( t_{i_2 \rightarrow L+1} \rightarrow 1 \). With \( u_{i_2 \rightarrow L+1} = 1 - t_{i_2 \rightarrow L+1}, c = \sqrt{2}\pi, \) and \( \prod_{i \in I} s_i := 1 \) if \( I \) is the empty set,

\[
K_{\text{NTK}}^{(L+1)} \left( \{1 - u_{i_2 \rightarrow L+1}\}_{i_2 \rightarrow L+1} \right) = L + 1 - \frac{c}{p_L} \sum_{l' = 1}^L \left( \prod_{l'' = l' + 1}^L s_{l''} \right) \sum_{i_{2 \rightarrow L+1}} \left( \prod_{2 \leq i'' \leq L} s_{i''} \right) \sum_{i_2 \rightarrow L+1} u_{i_2 \rightarrow L+1} \left( \frac{1}{2} \right)^{1/2} + O(u_{i_2 \rightarrow L+1}^{3/2}).
\]
Proposition C.4 (NTK when $x = - y$) The NTK of a hierarchical network of depth $L + 1$, filter sizes $(s_1, \ldots, s_L)$ and $p_l \geq 1$ has the following singular expansion when all $t_{i_{z=L+1}} \to -1$. With $u_{i_{z=L+1}} = 1 + t_{i_{z=L+1}}$, $c = \sqrt{2}/\pi$ and $\prod_{l \in I} s_l := 1$ if $I$ is the empty set,

$$
K_{\text{NTK}}^{(L+1)} \left( \left\{ 1 + u_{i_{z=L+1}} \right\}_{i_{z=L+1}} \right) = a_L + \frac{cL}{\prod_{l \in I} s_l} \sum_{l \leq i_{z=L+1}} u_l^{3/2} + O(u_l^{5/2}),
$$

(S33)

with $a_L = b_{L} + b_{L-1} \kappa_0(b_{L-1}), b_{L} = \kappa_1(b_{L-1}), b_{1} = 0$; and $c_L = c_{L-1} \kappa_0(b_{L-1}), c_{1} = c$. Notice that both $\kappa_1$ and $\kappa_0$ are positive and strictly increasing in $[0, 1]$ and $\kappa_1(1) = \kappa_0(1) = 1$, thus $b_{1} \in (0, 1)$ and $c_{L} < c_{L-1}$.

The proofs of the two propositions above are omitted, as they follow the exact same steps of the previous two proofs.

C.2 Patches on the ring

In this section, we prove a restricted version of Thm. 3.1 for the case of 2-dimensional input patches, since the reduction of spherical harmonics to the Fourier basis simplifies the proof significantly. We also consider, for convenience, hierarchical kernels of depth 3 with the filter size of the second hidden layer set to $p = d/2$, the total number of 2-patches of the input. Once this case is understood, extension to arbitrary filter size and arbitrary depth is trivial.

Theorem C.1 (Spectrum of depth-3 kernels on 2-patches) Let $T_K$ be the integral operator associated with a $d$-dimensional hierarchical kernel of depth 3, (2 hidden layers), with filter sizes $(s_1 = 2, s_2)$ and $p_2 = 1$, such that $2s_2 = d$ and $s_2 = p$ (the number of 2-patches). Eigenvalues and eigenfunctions of $T_K$ can be organised into 2 sectors associated with the hidden layers of the kernel/network.

i. The first sector consists of $s_1$-local eigenfunctions, which are functions of a single patch $x$, for $i = 1, \ldots, p$. The labels $k, \ell$ of local eigenfunctions are such that all the $k_i$’s with $j \neq i$ are zero (because the eigenfunction is constant outside $x$). The corresponding eigenvalue is degenerate with respect to the location of the patch: we call it $\Lambda^{(1)}$. When $k_i \to \infty,$

$$
\Lambda^{(1)} = c_{2, 1} k^{-2 \nu - 1} + o \left( k^{-2 \nu - 1} \right),
$$

with $\nu_{\text{NTK}} = 1/2, \nu_{\text{HFK}} = 3/2$. $C_{2, 1}$ can take two distinct strictly positive values depending on the parity of $k_i$;

ii. The second sector consists of global eigenfunctions, which are functions of the whole input $\mathbf{x}$. The labels $k, \ell$ of global eigenfunctions are such that at least two of the $k_i$’s are non-zero. We call the corresponding eigenvalues $\Lambda^{(2)}$. When $\|k\| \to \infty$, with $k = \|k\|$,

$$
\Lambda^{(2)} = c_{2, 2} k^{-2 \nu - p} + o \left( k^{-2 \nu - p} \right),
$$

Proof. If we consider binary patches in the first layer, the input space becomes the Cartesian product of two-dimensional unit spheres, i.e. circles, $\mathcal{X} = \prod_{i=1}^{d} S^1$. Then, each patch $x$ corresponds to an angle $\theta_i$ and the spherical harmonics are equivalent to Fourier atoms,

$$
Y_0(\theta) = 1, \quad Y_{k,1}(\theta) = e^{i k \theta}, \quad Y_{k,2}(\theta) = e^{-i k \theta}, \quad \forall k \geq 1.
$$

(S36)

Therefore, solving the eigenvalue problem for a dot-product kernel $K(x \cdot y) = K(\cos(\theta_x - \theta_y))$ with $x, y \in S^1$ reduces to computing its Fourier transform. With $|S^1| = 2$ and $|S^2| = 2\pi$,

$$
\frac{1}{2\pi} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} d\theta d\phi \ K(\cos(\theta_x - \theta_y)) e^{\pm i \kappa \theta_x} = \Lambda_k e^{\pm i \kappa \theta_x} \Rightarrow \Lambda_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} d\theta \ K(\cos \theta) e^{\pm i \kappa \theta},
$$

(S37)

where we denoted with $\theta$ the difference between the two angles. Similarly, for a multi-dot-product kernel, the eigenvalues coincide to the $p$-dimensional Fourier transform of the kernel, where $p$ is the
number of patches,
\[
\Lambda_k = \frac{1}{(2\pi)^p} \int_{-\pi}^{\pi} \left( \prod_{i=1}^{p} d\theta_i e^{\pm ik_i \theta_i} \right) K \left( \left\{ \cos \theta_i \right\}_{i=1}^{p} \right) = \frac{1}{(2\pi)^p} \int_{-\pi}^{\pi} d^p \theta e^{\pm ik \cdot \theta} K \left( \left\{ \cos \theta_i \right\}_{i=1}^{p} \right), \tag{S38}
\]
with \( k = (k_1, \ldots, k_p)^T \) the vector of the patch wavevectors and \( \theta = (\theta_1, \ldots, \theta_p)^T \) the vector of the shift angles.

The nonanalyticity of the kernel at \( t_i = 1 \) for all \( i \) moves to \( \theta_i = 0 \) for all \( i \), whereas those in \( t_i = -1 \) move to \( \theta_i = \pi \) and \( -\pi \). The corresponding singular expansion is obtained from Eq. S24 after replacing \( t_i \) with \( \cos (\theta_i) \) and expanding \( \cos (\theta_i) \) as \( 1 - \theta_i^2/2 \), resulting in
\[
\mathcal{K}(\cos \theta_i)_{n=1} = 1 - \frac{1}{2p} \sum_{i=1}^{p} \theta_i^2 + \frac{1}{3\pi^2} \sum_{i=1}^{p} |\theta_i|^3 + \frac{2\sqrt{2}}{3\pi} \left( \frac{1}{p} \sum_{i=1}^{p} \theta_i^2 \right)^{3/2} + O(\theta_i^4). \tag{S39}
\]
The first nonanalytic terms are \( \frac{1}{3\pi} \sum_{i=1}^{p} |\theta_i|^3 \) and \( \frac{2\sqrt{2}}{3\pi} \left( \frac{1}{p} \sum_{i=1}^{p} \theta_i^2 \right)^{3/2} \). After recalling that the Fourier transform of \( |\theta|^{2\nu} \) with \( \nu = 3/2 \) has \( (\nu = 3/2)\)
\[
\frac{1}{(2\pi)^p} \int_{-\pi}^{\pi} d^p \theta e^{\pm ik \cdot \theta} \sum_{i=1}^{p} |\theta_i|^3 \sim \sum_{i=1}^{p} k_i^{-4} \prod_{j \neq i} \delta_{k_j,0}, \quad \text{for } \|k\| \to \infty \tag{S40}
\]
and
\[
\frac{1}{(2\pi)^p} \int_{-\pi}^{\pi} d^p \theta e^{\pm ik \cdot \theta} \|\theta\|^{3} \sim \|k\|^{-p-3}, \quad \text{for } \|k\| \to \infty. \tag{S41}
\]
All the other terms in the kernel expansion will result to subleading contributions in the Fourier transform. Therefore, the former of the two equations above yields the asymptotic scaling of eigenvalues of the local sector, whereas the latter yields the asymptotic scaling of the global sector.

The proof for the NTK case is analogous to the RFK case, except that the singular expansion near \( \theta_i = 0 \) is given by
\[
\mathcal{N}TK(\cos \theta_i)_{n=1} = 3 - \frac{1}{p} \sum_{i=1}^{p} |\theta_i| - \sqrt{2} \frac{1}{\pi} \left( \frac{1}{p} \sum_{i=1}^{p} \theta_i^2 \right)^{1/2} + O(\theta_i^{3/2}). \tag{S42}
\]

### C.3 Patches on the \( s \)-dimensional hypersphere

In this section, we make an additional step towards Thm. 3.1 by extending Thm. C.1 to the case of \( s \)-dimensional input patches. We still consider hierarchical kernels of depth 3 with the filter size of the second hidden layer set to \( p = d/s \) (the total number of \( s \)-patches of the input) so as to ease the presentation. Extension to general depth and filter sizes is presented in Subsec. C.4.

**Theorem C.2 (Spectrum of depth-3 kernels on \( s \)-patches)** Let \( T_k \) be the integral operator associated with a \( d \)-dimensional hierarchical kernel of depth 3, (2 hidden layers), with filter sizes \( (s_1, s_2) \) and \( p_2 = 1 \), such that \( 2s_2 = d \) and \( s_2 = p \) (the number of \( s \)-patches). Eigenvalues and eigenfunctions of \( T_k \) can be organized into 2 sectors associated with the hidden layers of the kernel/network.

i. The first sector consists of \( s_1 \)-local eigenfunctions, which are functions of a single patch \( x_i \) for \( i = 1, \ldots, p \). The labels \( k, \ell \) of local eigenfunctions are such that all the \( k_j \)'s with \( j \neq i \) are zero (because the eigenfunction is constant outside of \( x_i \)). The corresponding eigenvalue is degenerate with respect to the location of the patch: we call it \( \Lambda^{(1)}_{k_i} \). When \( k_i \to \infty \),
\[
\Lambda^{(1)}_{k_i} = C_{s_1} k^{-2\nu - (s-1)} + o \left( k^{-2\nu - (s-1)} \right), \tag{S43}
\]
with \( \nu_{NTK} = 1/2, \nu_{RFK} = 3/2 \). \( C_{s_1} \) can take two distinct strictly positive values depending on the parity of \( k_i \);
ii. The second sector consists of global eigenfunctions, which are functions of the whole input \( x \). The labels \( k, \ell \) of global eigenfunctions are such that at least two of the \( k_i \)'s are non-zero. We call the corresponding eigenvalue \( \Lambda_k(2) \). When \( k = \|k\| \to \infty \), for fixed non-zero angles \( k/k_i \),

\[
\Lambda_k(2) = C_{s,2} \left( \frac{k}{k_i} \right) k^{-2\nu - p(s-1)} + o \left( k^{-2\nu - p(s-1)} \right),
\]

where \( C_{s,2} \) is a positive function.

**Proof.** A hierarchical RFK/NTK is a multi-dot-product kernel, therefore its eigenfunctions are products of spherical harmonics \( \hat{Y}_{k,\ell}(x) = \prod_{i=1}^p Y_{k_i,\ell_i}(x_i) \) and the eigenvalues of \( \mathcal{K} \) are given by [Eq. S47].

\[
\Lambda_k = \left( \prod_{i=1}^p \frac{1}{|S^{d-2}_s|} \right) \int_{-1}^{+1} dt_i \left( 1 - t_i^2 \right) \frac{\nu_s + \frac{d-3}{2}}{\Gamma(k_i + \frac{d-3}{2})} \mathcal{K} \left( \{t_i\} \right).
\]

The proof follows the following strategy: first we show that the infinitely differentiable part of \( \mathcal{K} \) results in eigenvalues which decay faster than any polynomial of the degrees \( k_i \). Then we show that the decay is controlled by the most singular term of the singular expansion of the kernel and finally compute such decay by relating it to the number of derivatives of the kernel having a finite \( l_2 \) norm.

When \( \mathcal{K} \) is infinitely differentiable in \([-1, +1]\), we can plug Rodrigues’ formula [Eq. S5] for each \( P_{k_i,s}(t_i) \) and get

\[
\Lambda_k = \left( \prod_{i=1}^p \frac{1}{|S^{d-2}_s|} \right) \int_{-1}^{+1} dt K(t) \left( \prod_{i=1}^p \frac{d^{k_i}}{dt_i^{k_i}} \left( 1 - t_i^2 \right)^{k_i + \frac{d-3}{2}} \right),
\]

with \( \int_{-1}^{+1} dt \) denoting integration over the \( p \)-dimensional hypercube \([-1, +1]^p\). We can simplify the integral further via integration by parts, so as to obtain

\[
\Lambda_k = \left( \prod_{i=1}^p \frac{1}{|S^{d-2}_s|} \right) \frac{1}{\Gamma(k_i + \frac{d-3}{2})} \int_{-1}^{+1} dt K^{(k)}(t) \left( \prod_{i=1}^p \left( 1 - t_i^2 \right)^{k_i + \frac{d-3}{2}} \right),
\]

where \( K^{(k)} \) denotes the partial derivative of order \( k_i \) with respect to \( t_1, t_2 \) with respect to \( t_2 \) and so on until \( k_p \) with respect to \( t_p \). Notice that the function \( (1 - t^2)^{\frac{d-3}{2}} \) is proportional to the probability measure of the scalar product \( t \) between two points sampled uniformly at random on the unit sphere [48 Sec. 1.3],

\[
|S^{d-1}| = \int_{-1}^{+1} dt \left( 1 - t^2 \right)^{\frac{d-3}{2}} \int_{d-2}^{d-1} dS^{d-2} \int_{-1}^{+1} dt \left( 1 - t^2 \right)^{\frac{d-3}{2}} = 1.
\]

This probability measure converges weakly to a Dirac mass \( \delta(t) \) when \( d \to \infty \). Recall, in addition, that \( |S^{d-1}| = 2\pi^{d/2}/\Gamma(d/2) \), where \( \Gamma \) denotes the Gamma function \( \Gamma(x) = \int_0^\infty dx x^{d-1} e^{-x} \). Thus, with converges weakly to a Dirac measure \( \delta(t) \) as \( c \to \infty \), once properly rescaled. In particular, choosing \( k_i \) such that \( k_i + (s-3)/2 = (d-3)/2 \), one has

\[
\lim_{k_i \to \infty} \frac{\Gamma \left( k_i + \frac{d}{2} \right)}{\sqrt{\pi} \Gamma \left( k_i + \frac{d-3}{2} \right)} \left( 1 - t_i^2 \right)^{k_i + \frac{d-3}{2}} = \delta(t_i).
\]

As a result, when \( \mathcal{K} \) is infinitely differentiable, one has the following equivalence in the limit where all \( k_i \)'s are large,

\[
\Lambda_k \sim \left( \prod_{i=1}^p \frac{1}{|S^{d-2}_s|} \right) \frac{1}{\Gamma(k_i + \frac{d-3}{2})} \frac{\nu_s + \frac{d-3}{2}}{\Gamma(k_i + \frac{d-3}{2})} \mathcal{K}^{(k)}(0),
\]

which implies that, when \( \mathcal{K} \) is infinitely differentiable, the eigenvalues decay exponentially or faster with the \( k_i \).

Let us now consider the nonanalytic part of \( \mathcal{K} \). There are three kinds of terms appearing in the singular expansion of depth-3 kernels (cf. Subsec. C.1):

\[\text{ia) } \sum_i (1 - t_i)^\nu \text{ near } t_i = +1;\]
where we have introduced $N_{k,s}$ as the number of $k$-th Legendre polynomial. The asymptotic decay of $(f_1)_k$ is strictly related to the differentiability of $f_1$, which is in turn controlled by action of the Laplace-Beltrami operator $\Delta$ on $f_1$. As a function on the sphere $S^{s-1}$, $f_1$ depends only on one angle, therefore the Laplace-Beltrami operator acts as follows,

$$\Delta f_1(\theta) = \frac{1}{\sin(\theta)^{s-2}} \frac{d}{d\theta} \left( \sin(\theta)^{s-2} f_1(\theta) \right) = f_1''(\theta) + (d-2) \frac{\cos(\theta)}{\sin(\theta)} f_1'(\theta).$$

In terms of singular behaviour near $\theta = 0$, $f_1(\theta) \sim |\theta|^{2\nu}$ implies $\Delta f_1(\theta) \sim |\theta|^{2\nu - 2}$, thus $\Delta^m f_1(\theta) \sim |\theta|^{2(\nu - m)}$. Given $\nu$, repeated applications of $\Delta$ eventually result in a function whose $L^2$ norm on the sphere diverges. On the one hand,

$$\|\Delta^{m/2} f_1\|^2 = \int_0^\pi \sin^{d-2}(\theta) f_1(\theta) \Delta^m f_1(\theta).$$

The integrand behaves as $|\theta|^{d-2+4\nu-2m}$ near 0, thus the integral diverges for $m \geq 2\nu + (d-1)/2$. On the other hand, from Eq. S.3,

$$\|\Delta^{m/2} f_1\|^2 = \sum_k N_{k,s} (k(2s-2))^m \|f_1\|^2.$$  

As $N_{k,s} \sim k^{s-2}$ and the sum must converge for $m < 2\nu + (d-1)/2$ and diverge otherwise, $(f_1)_k \sim k^{-2\nu - (s-1)}$. The projections of all the other terms in $K$ on Legendre polynomials of one of the $p$ angles $\theta_i$ display a faster decay with $k$, therefore the above results implies the asymptotic scaling of local eigenvalues. Notice that such scaling matches with the result of [25], which was obtained with a different argument.

Finally, let us look at the contribution to the eigenvalue $\Lambda_k$ due to the term $f_2(||\theta||)$:

$$\left( \prod_{j=1}^p \frac{|S^{s-2}|}{|S^{s-1}|} \int_0^\pi d\theta_j (\sin(\theta_j))^{s-2} P_{k,s}(\cos(\theta_j)) \right) f_2(||\theta||) = (f_2)_k,$$

where we have introduced $(f_2)_k$ as the projection of $f_2(||\theta||)$ on the multi-Legendre polynomial with multi-degree $k$. The asymptotic decay of $(f_2)_k$ is again related to the differentiability of $f_2$, controlled by action of the multi-sphere Laplace-Beltrami operator $\Delta_{p,s}$ in Eq. S.13. As $f_2$ depends only on one angle per sphere,

$$\Delta_{p,s} f_2(||\theta||) = \sum_{i=1}^p \left( \partial^2_{\theta_i} f_2(||\theta||) + (s-2) \frac{\cos(\theta_i)}{\sin(\theta_i)} \partial_{\theta_i} f_2(||\theta||) \right).$$

(S57)
Further simplifications occur since \( f_2 \) depends only on the norm of \( \theta \). In terms of the singular behaviour near \( \| \theta \| = 0 \), \( f_2 \sim \| \theta \|^{2\nu} \) implies \( \Delta^{m/2}_{p,s} f_2 \sim \| \theta \|^{2(\nu-m)} \), thus

\[
\| \Delta^{m/2}_{p,s} f_2 \|^2 = \int_{[0, \pi]^p} d\theta \prod_{i=1}^{p} (\sin^{s-2}(\theta_i)) f_2(\| \theta \|) \Delta^{m}_{p,s} f_2(\| \theta \|) < +\infty \tag{S58}
\]

requires \( m < 2\nu + p(s-1)/2 \) (compare with \( m < 2\nu + (s-1)/2 \) for the local contributions). Therefore, one has

\[
\| \Delta^{m/2}_{p,s} f_1 \|^2 = \sum_{k} \left( \prod_{i=1}^{p} \mathcal{N}_{k_i,s} \right) \left( \sum_{i=1}^{p} k_i (k_i + s - 2) \right)^m |(f_2)_k|^2 < +\infty \quad \forall m < 2\nu + p(s-1)/2,
\]

while the sum diverges for for \( m \geq 2\nu + p(s-1)/2 \). In addition, since \( f_2 \) is a radial function of \( \theta \) which is homogeneous (or scale-invariant) near \( \| \theta \| = 0 \), \( (f_2)_k \) can be factorised in the large-\( \| k \| \) limit into a power of the norm \( \| k \|^{\alpha} \) and a finite angular part \( C(k/\| k \|) \). By plugging the factorisation into Eq. S59 we get

\[
(f_2)_k \sim \mathcal{C}(k/\| k \|) \| k \|^{-2\nu-p(s-1)} \sum_{k/\| k \| = k} \left( \prod_{i=1}^{p} (k_i/k)^{s-2} \right) \mathcal{C}(k/\| k \|)^2 < +\infty \tag{S60}
\]

The projections of all the other terms in \( K \) on multi-Legendre polynomials display a faster decay with \( \| k \| \), therefore the above results implies the asymptotic scaling of global eigenvalues.

C 4 General depth

The generalisation to arbitrary depth is trivial once the depth-3 case is understood. For global and \( s_l \)-local eigenvalues, the analysis of the previous section carries over unaltered. All the other intermediate sectors correspond to the other terms singular expansion of the kernel: from Subsec. C.1 these terms can be written as

\[
\frac{c}{p! \prod_{l' < l'' \leq L} s_{l''}} \sum_{\prod_{l' < l'' \leq L} s_{l''}} \left( \prod_{2 \leq l' < l''} \frac{1}{s_{l''}} \right) \sum_{\prod_{l' < l'' \leq L} s_{l''}} (1 - t_{i_2 \rightarrow L+1})^{\nu}, \tag{S61}
\]

for some \( l'' = 2, \ldots, L - 1 \) and fractional \( \nu \). In practice, this term is a sum over the \( p_{l''} = p! \prod_{l' < l'' \leq L} s_{l''} \) meta-patches of \( t \) having size \( s_{2 \rightarrow l''} := \prod_{2 \leq l' < l''} s_{l''} \). Each summand is the fractional power \( \nu \) of the average of the \( t_i \)’s within a meta-patch. When plugging such term into Eq. S45 the integrals over the \( t_i \)’s which do not belong to that meta-patch yield Kronecker deltas for the corresponding \( k_i \)’s. The integrals over the \( t_i \)’s within the meta-patch, instead, can be written as in Eq. S56 with the product and the norm restricted over the elements of that meta-patch, i.e., \( \| \theta \| \rightarrow \left( \sum_{t_{i_2 \rightarrow L+1}} \theta_{i_2 \rightarrow L+1}^2 \right)^{1/2} \). Therefore, the scaling of the eigenvalue with \( k \) is given again by Eq. S61 but with \( p \) replaced by the size of the meta-patch \( \prod_{2 \leq l' < l''} s_{l''} \), so that the effective dimension of Eq. 3 appears at the exponent.

D Generalisation bounds for kernel regression and spatial adaptivity

Consider the regression setting detailed in Sec. 4 of the main text. First, assume that the target function \( f^* \) belongs to the KKHS \( \mathcal{H} \) of the kernel \( K \). Then, without further assumptions on \( K \), we have the following dimension-free bound on the excess risk, based on Rademacher complexity [33 Chs. 4, 7], [13],

\[
\mathcal{E}(\lambda, n) - \epsilon(f^*) \leq C \| f^* \|_{\mathcal{H}} \sqrt{\frac{\mathrm{Tr}(\mathcal{T}_K)}{n}}, \tag{S62}
\]

where \( \mathcal{T}_K \) is the integral operator associated to \( K \). For a hierarchical kernel, having a target with more power in the local sectors can result in a smaller \( \| f^* \|_{\mathcal{H}} \), hence smaller excess risk. However,
this gain is only a constant factor in terms of sample complexity and, more importantly, being in the RKHS requires an order of smoothness which typically is of the order of the dimension, which is a very-restrictive assumption in high-dimensional settings. This result can be extended by including

Finally, when \( r > 1 \), the bound becomes

In short, the replica calculation used to obtain these equations consists in defining an energy functional \( \mathcal{E}(f) \) related to the empirical MSE and assigning to the predictor \( f \) a Boltzmann measure, i.e. \( P(f) \sim e^{-\beta \mathcal{E}(f)} \). When \( \beta \to \infty \), the measure concentrates around the minimum of \( \mathcal{E}(f) \), which coincides with the minimiser of the empirical MSE. Then, since \( \mathcal{E}(f) \) depends only quadratically on the projections \( c_\rho \), computing the average over data that appears in the definition of the generalisation error, reduces to computing Gaussian integrals. While non-rigorous, this method has been successfully used in physics—to study disordered systems—and in machine learning theory. In particular, the
predictions obtained with Eq. S68 and Eq. S69 has been validated numerically for both synthetic and real datasets. In Eq. S68 \( \kappa_\lambda(n)/n \) plays the role of a threshold: the modal contributions to the error tend to 0 for \( \rho \) such that \( \Lambda_\rho \gg \kappa_\lambda(n)/n \), and to \( \mathbb{E}[e^2] \) for \( \rho \) such that \( \Lambda_\rho \ll \kappa_\lambda(n)/n \). This is equivalent to saying that kernel regression can capture only the modes corresponding to the eigenvalues larger than \( \kappa_\lambda(n)/n \) (see also [52, 53]).

In the ridgeless limit \( \lambda \rightarrow 0^+ \), this threshold asymptotically tends to the \( n \)-th eigenvalue of the student, resulting in the intuitive picture presented in the main text. Namely, given \( n \) training points, ridgeless regression learns the \( n \) projections corresponding to the highest eigenvalues. In particular, assume that the kernel spectrum and the target function projections decay as power laws. Namely, \( \Lambda_\rho \sim \rho^{-a} \) and \( \mathbb{E}[c_\rho^2] \sim \rho^{-b} \), with \( 2a > b - 1 \). Furthermore, we can approximate the summations over modes with an integral by using the Euler-MacLaurin formula. Hence, we substitute the eigenvalues with their asymptotic limit \( \Lambda_\rho = A \rho^{-a} \). Since, \( \kappa_0(n)/n \rightarrow 0 \) as \( n \rightarrow \infty \), these two operations result in an error which is asymptotically independent of \( n \). In particular,

\[
\frac{\kappa_0(n)}{n} = \frac{\kappa_0(n)}{n} \left( \int_0^\infty \frac{d\rho \ A \rho^{-a}}{A \rho^{-a} + \kappa_0(n)/n} + O(1) \right)
= \frac{\kappa_0(n)}{n} \left( \kappa_0(n) \right) \left( \frac{1}{\rho} \right) \int_0^\infty d\sigma \sigma^{\frac{1}{a}-1} A^{\frac{1}{a}-1} + O(1) \right).
\]

(S70)

Since the integration over \( \sigma \) is finite and independent of \( n \), we obtain that \( \kappa_0(n)/n = O(n^{-a}) \).

Similarly, we find that the mode-independent prefactor \( \partial_\lambda (\kappa_\lambda(n)/n) |_{\lambda=0} = O(1) \).

As a result, we have

\[
\epsilon(n) \sim \sum_{\rho} \frac{n^{-2a}}{A \rho^{-a} + n^{-a}} \mathbb{E}[e^2].
\]

(S71)

Following the intuitive argument about the thresholding action of \( \kappa_0(n)/n \sim n^{-a} \), we can split the summation in Eq. S71 into modes where \( \Lambda_\rho \gg \kappa_0(P)/n \), \( \Lambda_\rho \sim \kappa_0(n)/n \) and \( \Lambda_\rho \ll \kappa_0(n)/n \),

\[
\epsilon(n) \sim \sum_{\rho \ll n} \frac{n^{-2a}}{(A \rho^{-a})} \mathbb{E}[e^2] + \sum_{\rho \approx n} \frac{1}{2} \mathbb{E}[e^2] + \sum_{\rho \gg n} \mathbb{E}[e^2].
\]

(S72)

Finally, Eq. 18 is obtained by noticing that, under the assumption on the decay of \( \mathbb{E}[c^2_\rho] \), the contribution of the summation over \( \rho \ll n \) is subleading in \( n \), whereas the other two can be merged together.

### F Learning curves of teacher-student scenarios

#### F.1 Recap of error rates

In this section, we recall and comment on the results obtained in the different teacher-student scenarios. To ease notation, in the following we always consider the NTK, i.e. smoothness exponent \( \nu = 1/2 \).

The first case consists of one-hidden-layer convolutional teacher (left) and student (right) kernels. As highlighted in blue, the output of the teacher is a linear combination (dashed lines indicate the linear output weights) of \( s_1 \)-dimensional functions of the input patches. If the structure of the student is matched to the one of the teacher, the learning problem becomes effectively \( (s_1 - 1) \)-dimensional and the error decays as \( n^{-1/(s_1 - 1)} \), instead of \( n^{-1/d_{\text{eff}}} \), with \( d_{\text{eff}} \) the total input dimension with

\[
\tau(n) \sim n^{-\frac{1}{s_1-1}}
\]

As highlighted in blue, the output of the teacher is a linear combination (dashed lines indicate the linear output weights) of \( s_1 \)-dimensional functions of the input patches. If the structure of the student is matched to the one of the teacher, the learning problem becomes effectively \( (s_1 - 1) \)-dimensional and the error decays as \( n^{-1/(s_1 - 1)} \), instead of \( n^{-1/d_{\text{eff}}} \), with \( d_{\text{eff}} \) the total input dimension with
the number of spherical constraints subtracted (one per patch). Notice that the role of the student’s structure, i.e. the algorithm, is as crucial as the role of the teacher, i.e. the task. Indeed, using a fully-connected student with no prior on the task’s locality would result in an error’s decay cursed by dimensionality. However, in contrast to fully-connected students, shallow convolutional students are only able to learn tasks with the same structure. In particular, any task entailing non-linear interactions between patches—which are arguably crucial in order to learn image data—belongs to their null space.

As we illustrated in the main text, to solve this strong constraint on the hypothesis space, one has to consider deep convolutional architectures. In particular, consider the same shallow teacher of the previous paragraph (left) learnt by a depth-four convolutional student (right).

Remarkably, this student is able to learn the teacher without being cursed by input dimensionality. Indeed, as the number of patches diverges, the error decay asymptotes to $n^{-\frac{1}{s_1}}$. This rate is slightly worse than the one obtained by the student matched with the teacher, which is proven to be the Bayes-optimal case, but far from being cursed. Intuitively, this fast rate is obtained because the student eigenfunctions of the first sector, i.e. constant outside a single patch, correspond to large eigenvalues and bias the learning dynamics towards $s_1$-local functions. Yet, this student is also able to represent functions which are considerably more complex.

Consider a depth-three teacher (left) learned by a depth-four student (right).

As highlighted in orange, the output of the teacher is a linear combination of a composition of non-linear functions acting on patches and coupling them. In this setting, the error decay is controlled by the effective dimension of the second layer. In fact, when the number of patches diverges, the error decay asymptotes to $n^{-1/d_{eff}(2)}$. In general, this behaviour is a result of what we called ‘adaptivity to the spatial structure’ of the target.

Finally, consider both teacher and student with the complete hierarchy, i.e. the receptive fields of the neurons in the penultimate layers coincide with the full input.
In this case, we show that the error decays as \( n^{-1/d_{alt}(3)} \), i.e. the rate is cursed by the input dimension. The physical meaning of this result is that the hierarchical structure we are considering is still too complex and cannot be learnt efficiently. In other words, these hierarchical convolutional networks are excellent students, since they are able to adapt to the spatial structure of the task, but bad teachers, since they generate global functions which are too complex to be learnt efficiently.

**F.2 Shallow architectures**

Consider depth-two teacher and student kernels with filters of size \( s \). Our goal is to estimate the asymptotic decay of the average-case generalisation error using the spectral bias formulation,

\[
\tau(n) \sim \sum_{k \text{ s.t. } \Lambda_k < \Lambda^S(n)} \text{E}_{f^*} \left[ (f_{k, t}^*)^2 \right]. \tag{S73}
\]

First, we compute the scaling of the \( n \)-th eigenvalue of the student kernel \( \Lambda_{k,t} \) with the number of samples \( n \). Since \( L = 1 \), the student eigenvalues decay as

\[
\Lambda_k^S \sim \sum_{i=1}^{p} k_i^{-2\nu_S-(s-1)} \prod_{j \neq i} \delta_{k_j,0}. \tag{S74}
\]

In order to take into account their algebraic multiplicity, we introduce the eigenvalue density \( D^S(\Lambda) \), whose asymptotic form for small eigenvalues is

\[
D^S(\Lambda) = \sum_{k,t} \delta(\Lambda - \Lambda_k^S) \\
\sim \sum_{k} \left( \prod_{i=1}^{p} k_i^{s-2} \right) \delta \left( \Lambda - \sum_{i=1}^{p} k_i^{-2\nu_S-(s-1)} \prod_{j \neq i} \delta_{k_j,0} \right) \\
\sim \sum_{i=1}^{p} \sum_{k_i} k_i^{s-2} \delta \left( \Lambda - k_i^{-2\nu_S-(s-1)} \right) \\
\sim \int_{1}^{\infty} dk k^{s-2} \delta \left( \Lambda - k^{-2\nu_S-(s-1)} \right) \\
\sim \Lambda^{-1-\frac{s-1}{2\nu_S+(s-1)}}. \tag{S75}
\]

Thus, the scaling of \( \Lambda^S(n) \) can be determined self-consistently,

\[
n = \int_{\Lambda^S(1)}^{\Lambda^S(n)} d\Lambda D(\Lambda) \sim \Lambda^S(n)^{-\frac{s-1}{2\nu_S+(s-1)}} \Rightarrow \Lambda^S(n) \sim n^{-\frac{2\nu_S}{s-1}}. \tag{S76}
\]

Therefore, the set \( \{ k \text{ s.t. } \Lambda_k^S < \Lambda^S(n) \} \) is the set of local \( k_i \)'s larger than \( n^{1/(s-1)} \).

Then, since \( i \) the target is a Gaussian random field with the teacher kernel \( K_T \) as covariance, and \( ii \) teacher and student kernels share the same eigenfunctions, the target’s squared projection on the student eigenbasis are simply the eigenvalues of \( K_T \). Namely,

\[
\text{E}_{f^*} \left[ f_{k,t}^* f_{k',t}^* \right] = \Lambda_k^T \delta_{k,k'} \delta_{t,t'} \Rightarrow \text{E}_{f^*} \left[ (f_{k,t}^*)^2 \right] = \Lambda_k^T. \tag{S77}
\]

Finally, plugging everything in Eq. 73

\[
\tau(n) \sim \sum_{k_i > n^{1/(s-1)}} \sum_{k_i} k_i^{-2\nu_S-(s-1)} \\
\sim \int_{n^{1/(s-1)}}^{\infty} dk k^{s-2} k^{-2\nu_S-(s-1)} \\
\sim n^{-\frac{2\nu_S}{s-1}}. \tag{S78}
\]
F.3 Deep architectures

Consider the general setting with a student of depth \( L + 1 \) with filter sizes \((s_1, \ldots, s_L)\) and \( p_L = 1\), and a teacher of depth \( l + 1, l \leq L\), with filter sizes equal to those of the student up to \( s_l\). For each sector \( l\), one can compute the density of eigenvalues \( D_{(l)}(\Lambda)\). Depending on \( s_1\), there are two different cases.

If \( s_1 = 2\),

\[
D_{(l)}^{S}(\Lambda) = \sum_{\ell, k} \delta(\Lambda - \Lambda_{k}^{(l)}) \\
\sim \sum_{i_{l}+1 \rightarrow L+1} \sum_{k_{i_{l}+1 \rightarrow L+1}} \delta \left( \Lambda - C_{2,l} ||k_{i_{l}+1 \rightarrow L+1}||^{-2\nu_{S} - d_{\text{eff}}(l)} \right) \\
\sim \int_{1}^{\infty} dk \, k^{d_{\text{eff}}(l)-1} \delta \left( \Lambda - C_{2,l} k^{-2\nu_{S} - d_{\text{eff}}(l)} \right) \\
\sim \Lambda^{-1} \frac{d_{\text{eff}}(l)}{2\nu_{S} + d_{\text{eff}}(l)}. \quad (S79)
\]

In order to generalise the result above to \( s_1 \geq 3 \) one must assume that the scaling of the density of eigenvalues is controlled by the eigenvalues whose degree labels \( k_{i} \) are all equal, so that the scaling of \( \text{Thm. 3.1} \) can be used. In other words, when the non-vanishing \( k_{i}'s \) are all large but different, then the eigenvalue must decay faster than \( ||k||^{-2\nu_{S} - d_{\text{eff}}(l)} \). Then

\[
D_{(l)}^{S}(\Lambda) = \sum_{\ell, k} \delta(\Lambda - \Lambda_{k}^{(l)}) \\
\sim \sum_{i_{l}+1 \rightarrow L+1} \sum_{k_{i_{l}+1 \rightarrow L+1}} \delta \left( \Lambda - C_{s_{1},l} k_{i_{l}+1 \rightarrow L+1}^{-2\nu_{S} - d_{\text{eff}}(l)} \right) \\
\sim \Lambda^{-1} \frac{d_{\text{eff}}(l)}{2\nu_{S} + d_{\text{eff}}(l)}. \quad (S80)
\]

When summing over all layers \( l's\), the asymptotic behaviour of the total density of eigenvalues \( D^{S}(\Lambda) = \sum_{l} D_{(l)}^{S}(\Lambda) \) is dictated by the density of the sector with the slowest decay, i.e. the last one. Hence,

\[
D^{S}(\Lambda) \sim \Lambda^{-1} \frac{d_{\text{eff}}(L)}{2\nu_{S} + d_{\text{eff}}(L)}. \quad (S81)
\]

Therefore, similarly to the shallow case, one finds self-consistently that the \( n\)-th eigenvalue of the student \( \Lambda^{S}(n) \) decays as

\[
\Lambda^{S}(n) \sim n^{-1} \frac{2\nu_{S}}{\nu_{T} + d_{\text{eff}}(T)}. \quad (S82)
\]

Since by construction the target function belongs only to the first \( l \) sectors, the only non-zero projections will be the ones on the vectors \( k \) belonging to these sectors. Thus, all the \( k's \) in the sectors of the layers \( l' > l \) do not contribute to the sum. In particular, the sum is again dominated by the \( k's \) of the largest sector and the set \( \{ k \ s.t. \ \Lambda_{k}^{S} < \Lambda^{S}(n) \} \) is the set of \( k_{i_{l}+1 \rightarrow L+1}'s \) with components larger than \( n^{-1} \frac{2\nu_{S}}{\nu_{T} + d_{\text{eff}}(T)} \). Plugging into Eq. S73 for \( s_1 = 2 \) we find

\[
\tau(n) \sim n^{-\frac{2\nu_{S}}{\nu_{T} + d_{\text{eff}}(T)}} \frac{2\nu_{S} + d_{\text{eff}}(l)}{\nu_{T} + d_{\text{eff}}(T)}, \quad (S83)
\]

and for \( s_1 \geq 3\),

\[
\tau(n) \sim n^{-\frac{2\nu_{S}}{\nu_{T} + d_{\text{eff}}(T)}} \frac{2\nu_{S} + d_{\text{eff}}(l)}{\nu_{T} + d_{\text{eff}}(T)}, \quad (S84)
\]

Our numerical tests confirm these results even for \( s_1 \geq 3\).

F.4 Source-capacity conditions

In this section, we compute the two exponents controlling the source-capacity conditions in the teacher-student setting. We recall that these two conditions are satisfied if one finds \( \alpha \geq 1 \) and
The first condition is always satisfied when the student eigenvalue decays as $\Lambda^S_{k} \sim \rho^{-\alpha}$, i.e.

$$\alpha = 1 + 2\nu_S/d_{\text{eff}}(L).$$  \hfill (S86)

Following [Cor. 4.1] the source condition only depends on the first $l$ sectors. Moreover, substituting the average squared projections with the eigenvalues of the covariance of the target Gaussian random field, i.e. the teacher kernel, one finds

$$\left\| T_{K_S}^{1/\alpha} f^T \right\|_H^2 = \sum_{l' = 1}^l \sum_{i_{l'}, j_{l' + 1} \to L_{l'}} \left( \Lambda_{k_{i_{l'}, j_{l' + 1} \to L_{l'}}}^{S(l')} \right)^{-r} \Lambda_{k_{i_{l'}, j_{l' + 1} \to L_{l'}}}^{T(l')} .$$ \hfill (S87)

Plugging the relevant decays, for $s_1 = 2$,

$$\left\| T_{K_S}^{1/\alpha} f^T \right\|_H^2 = \sum_{l' = 1}^l \sum_{i_{l'}, j_{l' + 1} \to L_{l'}} \left( c_{2, l'} \| k_{i_{l'}, j_{l' + 1} \to L_{l'}} \|^{-2\nu_S - d_{\text{eff}}(l')} \right)^{-r} \times$$

$$\times c_{2, l'} \| k_{i_{l'}, j_{l' + 1} \to L_{l'}} \|^{-2\nu_T - d_{\text{eff}}(l')} .$$ \hfill (S88)

Since the most stringent condition to $r$ is given by the sector with the slowest eigenvalue decay, i.e. $l' = l$, in order to have a convergent series one needs

$$r < \frac{2\nu_T}{2\nu_S + d_{\text{eff}}(l)} .$$ \hfill (S89)

For $s_1 \geq 3$, in order to obtain the same result one needs to make the further assumption that when the non-vanishing $k_i$’s are all large but different, the eigenvalue must decay faster than $\| k \|^{-2\nu_T - d_{\text{eff}}(l')} .$

### G Numerical experiments

#### G.1 Experimental setup

Experiments were run on a high-performance computing cluster with nodes having Intel Xeon Gold processors with 20 cores and 192 GB of DDR4 RAM. All codes are written in PyTorch [54]. The repository containing all codes used to obtain the reported results can be found at [link temporarily removed for peer review].

#### G.2 Teacher-student learning curves

In order to obtain the learning curves, we generate $n + n_{\text{test}}$ random points uniformly distributed on the product of hyperspheres over the patches. We use $n \in \{128, 256, 512, 1024, 2048, 4096, 8192 \}$ and $n_{\text{test}} = 8192$. For each value of $n$, we sample a Gaussian random field with zero mean and covariance given by the teacher kernel. Then, we compute the kernel regression predictor of the student kernel, and we estimate the generalisation error as the mean squared error of the obtained predictor on the $n_{\text{test}}$ unseen example. The expectation over the teacher randomness is obtained by averaging over 16 independent sets of random input points and realisations of the Gaussian random fields. As teacher and student kernels, we use the analytical forms of the neural tangent kernels of hierarchical convolutional networks, with different combinations of depths and filter sizes.
Figure S1: Learning curves for deep convolutional NTKs ($\nu = 1/2$) in a teacher-student setting. **a.** Depth-two teachers learned by depth-two (matched) and depth-three (mismatched) students. Both these students are not cursed by the input dimension. **b.** Depth-three students learning depth-two and depth-three teachers. These students are cursed only in the second case. The numbers inside brackets are the sequence of filter sizes of the kernels. Solid lines are the results of experiments averaged over 16 realisations with the shaded areas representing the empirical standard deviations. The predicted asymptotic scaling $\epsilon \sim n^{-\beta}$ are reported as dashed lines.

**Depth-two and depth-three architectures.** [Fig. S1] reports the learning curves of depth-two and depth-three kernels with binary filters at all layers. Depth-three students defeat the curse of dimensionality when learning depth-two teachers, achieving a similar performance of depth-two students matched to the teacher’s structure. However, as we predict, these students encounter the curse of dimensionality when learning depth-three teachers.

**Ternary filters.** [Fig. S2] reports the learning curves for kernels with 3-dimensional filters and confirms our predictions in the $s_1 \geq 3$ case.

**Comparison with the noisy and optimally-regularised case.** Panel (a) of [Fig. S3] compares the learning curves obtained in the optimally-regularised and ridgeless cases for noisy and noiseless data, respectively. The first case corresponds to the setting studied in [22], in which the source-capacity formalism applies. In contrast with the second setting—which is the one used in the teacher-student scenarios and where it holds the correspondence between kernel methods and neural networks—i) we add to the labels a Gaussian random noise with standard deviation $\sigma = 0.1$, ii) for each $n$, we select the ridge resulting in the best generalisation performance. We observe that the decay obtained in the bound derived from the source-capacity conditions is exactly the one found numerically, i.e. the rate of the bound is tight. As a further check, panel (b) shows that the optimal ridge decays as prescribed.

**G.3 CIFAR-2 learning curves**

[Fig. S4] shows the learning curves of the neural tangent kernels of different architectures applied to pairs of classes of the CIFAR-10 dataset. In particular, the task is built by selecting two CIFAR-10 classes, e.g. plane and car, and assigning label $+1$ to the elements belonging to one class and label $-1$ to the remaining ones. Learning is again achieved by minimising the empirical mean squared error using a ‘student’ kernel. We find that the kernels with the worst performance are the ones corresponding to shallow fully-connected and convolutional architectures. Instead, for all the pairs of classes considered here, deep hierarchical convolutional kernels achieve the best performance.
Figure S2: Learning curves for deep convolutional NTKs ($\nu = 1/2$) with filters of size 3 in a teacher-student setting. a. Depth-three students learning depth-two and depth-three teachers. These students are cursed only in the second case. b. Depth-three models cursed by the effective input dimensionality. The numbers inside brackets are the sequence of filter sizes of the kernels. Solid lines are the results of experiments averaged over 16 realisations with the shaded areas representing the empirical standard deviations. The predicted asymptotic scaling $\varepsilon \sim n^{-\beta}$ are reported as dashed lines.

Figure S3: Noisy (optimally-regularised) vs noiseless (ridgeless) learning curves for depth-three deep convolutional NTKs ($\nu = 1/2$) in a teacher-student setting. a. Comparison between the learning curves in the noisy and noiseless case. Dashed lines represent the rates predicted with source-capacity bounds and replica calculations, respectively. Shaded areas represent the empirical standard deviations. b. Decay of the optimal ridge with the number of training points.

Figure S4: Learning curves of the neural tangent kernels of fully-connected (F-NTK) and convolutional (C-NTK) networks with various depths learning to classify two CIFAR-10 classes in a regression setting. Deep hierarchical convolutional kernels achieve the best performance. Shaded areas represent the empirical standard deviations obtained averaging over different training sets. a. Plane vs car. b. Cat vs bird.