How Wide Convolutional Neural Networks Learn Hierarchical Tasks

Francesco Cagnetta [1] * Alessandro Favero [1] * Matthieu Wyart [1]

[1]EPFL, *Equal contribution.

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 harvest the compositional and hierarchical structure of natural data such as images. Yet, we lack a quantitative understanding of how such structure affects performance, 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 nature, the data generated by deep CNNs are too rich to be efficiently learnable in high dimensions.

How can CNNs learn in high-dimensions?

- Learning a generic high-dimensional function is plagued by the **curse of dimensionality**: the rate at which the generalisation error ϵ decays with the number of training examples n vanishes when the input dimension d grows, i.e. $\epsilon = O(n^{-\beta})$ with $\beta = 1/d$ [1].
- Convolutional neural networks (CNNs) can learn data whose dimensions can be in the hundreds or more. This points to the existence of some underlying structure that CNNs can leverage. What is this structure?
- A popular hypothesis is that learnable tasks are compositional and hierarchical: features at any scale are made of sub-features at smaller scales. Can we quantify the effect of this structure on the generalisation error's decay β?

Deep CNNs and the kernel regime

L-hidden-layers hierarchical CNN. Denote with σ the normalised ReLU function, $\sigma(x) = \sqrt{2}\max(0, x)$. For each input $\boldsymbol{x} \in \mathbb{R}^d$, the output of a L-hidden-layers hierarchical neural network can be defined recursively as follows.

$$f^{(1)}_{h,i}(\boldsymbol{x}) = \sigma\left(\boldsymbol{w}^{(1)}_h \cdot \boldsymbol{x}_i\right), \forall h \in [1 \,..\, H_1], \forall i \in [1 \,..\, p_1];$$

$$f^{(l)}_{h,i}(\boldsymbol{x}) = \sigma\left(\frac{1}{\sqrt{H_{l-1}}}\sum_{h'}\frac{\boldsymbol{w}^{(l)}_{h,h'} \cdot \left(\boldsymbol{f}^{(l-1)}_{h'}\right)_i}{\sqrt{s_l}}\right), \forall h \in [1 \,..\, H_l], i \in [1 \,..\, p_l], l \in [2 \,..\, L];$$

$$f(\boldsymbol{x}) = f^{(L+1)}(\boldsymbol{x}) = \frac{1}{\sqrt{H_L}}\sum_{h=1}^{H_L}\sum_{i=1}^{p_L}\frac{w^{(L+1)}_{h,i} f^{(L)}_{h,i}(\boldsymbol{x})}{\sqrt{p_L}}. \tag{1}$$

H_l denotes the width of the l-th layer, s_l the filter size (with $s_1 \equiv s$, the size of the input patches), p_l the number of patches (with $p_1 \equiv p$, the number of input patches). The p_l's and s_l's satisfy $s_1 p_1 = d$ and $p_{l-1} = s_l p_l$. $\boldsymbol{w}^{(1)}_h \in \mathbb{R}^{s_1}$, $\boldsymbol{w}^{(l)}_{h,h'} \in \mathbb{R}^{s_l}$, $w^{(L+1)}_{h,i} \in \mathbb{R}$.

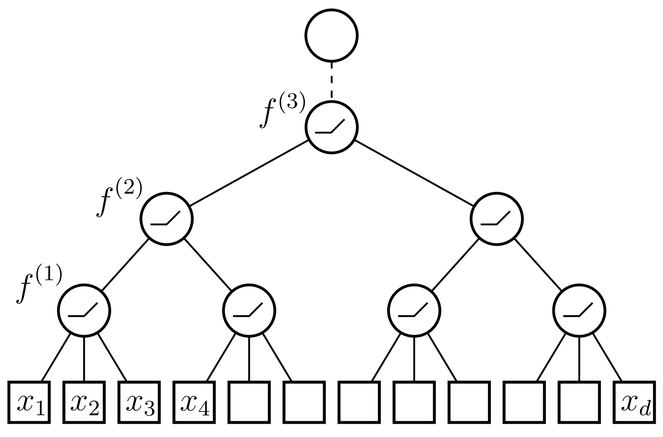

Figure 1. DAG of a hierarchical CNN with $L = 4$. The leaves represent the input coordinates, the root the output and all other nodes (channels of) hidden neurons.

Convolutional neural tangent kernel (CNTK). Assume that the weights of the network are initialised from a Gaussian distribution $\mathcal{N}(0, 1)$ and the number of channels diverges, i.e. $H_l \to \infty$. Then, denoting with $\boldsymbol{\theta}$ the collection of the network's parameter, the network's function $f(\boldsymbol{x}; \boldsymbol{\theta})$ learned by gradient descent on a square loss to zero training error is equivalent to the function obtained from ridgeless kernel regression with the neural tangent kernel (NTK) [2, 3]:

$$\mathcal{K}_{\text{NTK}}(\boldsymbol{x}, \boldsymbol{x}') = \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}; \boldsymbol{\theta}_0) \cdot \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}'; \boldsymbol{\theta}_0), \tag{2}$$

where $\boldsymbol{\theta}_0$ denotes the values of the parameters at initialisation. Interestingly, the NTK inherits the hierarchical structure of the network's function. In this limit, studying the generalisation properties of CNNs becomes equivalent to studying those of their NTK.

Statistical properties and spatial adaptation

Supervised learning setup. Consider a data distribution ρ on data-label pairs $(\boldsymbol{x}, y)$ such that $\mathbb{E}[y|\boldsymbol{x}] = f^*(\boldsymbol{x})$ for a target function f^*. Given n examples i.i.d. from ρ, the KRR estimator is given by

$$f^n_\lambda(\boldsymbol{x}) = \arg\min_{f \in \mathcal{H}} \frac{1}{n}\sum_{\mu=1}^{n}(f(\boldsymbol{x}_\mu) - y_\mu)^2 + \lambda\|f\|^2_{\mathcal{H}}, \tag{3}$$

where $\mathcal{H}$ is the RKHS of the kernel $\mathcal{K}$. The generalisation error of this estimator is defined as

$$\epsilon(f^n_\lambda) = \mathbb{E}_{(\boldsymbol{x},y)\sim\rho}[(f^n_\lambda(\boldsymbol{x}) - y)^2]. \tag{4}$$

Bound on the excess risk. The statistical properties of a kernel $\mathcal{K}$ are deeply connected with the eigendecomposition of its integral operator

$$\mathcal{T}_{\mathcal{K}} f(\boldsymbol{x}) = \int \mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') f(\boldsymbol{x}') d\tau(\boldsymbol{x}'), \tag{5}$$

with $d\tau$ the probability measure that generates the inputs. Let Λ_ρ denote its eigenvalues. if one can find $\alpha \geq 1$ and $r \geq 1 - 1/\alpha$ satisfying $\text{Tr}\left(\mathcal{T}^{1/\alpha}_{\mathcal{K}}\right) < \infty$ (capacity condition) and $\|T^{\frac{1-r}{2}}_{\mathcal{K}} f^*\|^2_{\mathcal{H}} < \infty$ (source condition), then, by choosing a n-dependent optimal λ_n, one gets the following bound on the excess risk [4, 5]:

$$\mathbb{E}_n[\epsilon(f^n_\lambda)] - \epsilon(f^*) \leq C\, n^{-\frac{\alpha r}{\alpha r + 1}}. \tag{6}$$

Spectral analysis of the CNTK. Let $d\tau$ be the uniform measure on the product of p s-dimensional unit spheres $\prod_i \mathbb{S}^{s-1}$. Then, the eigenfunctions of $\mathcal{T}_{\mathcal{K}_{\text{NTK}}}$ are $\tilde{Y}_{\boldsymbol{k},\boldsymbol{\ell}} = \prod_i Y_{k_i,\ell_i}(\boldsymbol{x}_i)$, where $Y_{k,\ell}$ denotes a spherical harmonic. In the paper, we show that the eigenfunctions of $\mathcal{T}_{\mathcal{K}_{\text{NTK}}}$ can be organised into groups associated with the hidden layers of the network. The eigenfunctions of each group have an *effective dimensionality* d_{eff}, which coincides with the receptive field of the neurons of the associated hidden layer. The corresponding eigenvalues scale as a power of the polynomial degree k of the eigenfunctions, with an exponent depending on d_{eff}. For $s = 2$, spherical harmonics reduce to Fourier atoms and the eigenfunctions reduce to the multidimensional Fourier basis. Then, for $\|\boldsymbol{k}\| \to \infty$,

$$\Lambda^{(l)}_{\boldsymbol{k}} = C_{2,l}\|\boldsymbol{k}\|^{-1-d_{\text{eff}}(l)} + o(\|\boldsymbol{k}\|^{-1-d_{\text{eff}}(l)}), \tag{7}$$

with $d_{\text{eff}} = s_2 s_3 \cdots s_l$.

Adaptivity to spatial structure.

(Informal) If the support of a target function f^* is entirely contained in a set of contiguous patches with dimension $d_{\text{eff}}(l)$ ($l \leq L$), then only the first l groups of the spectrum of $\mathcal{T}_{\mathcal{K}_{\text{NTK}}}$ contribute to the source condition. E.g., if f^* is Lipschitz, one obtains a fast rate of

$$n^{-2/(3+d_{\text{eff}}(l))} \tag{8}$$

as $d_{\text{eff}}(L) \to \infty$, which does not vanish with the input dimensionality!

Statistical mechanics of generalisation for the ridgeless limit. In the ridgeless case, where the correspondence between kernel methods and neural networks actually holds, there are no rigorous bounds. However, one can derive similar results using the replica methods of statistical physics [6].

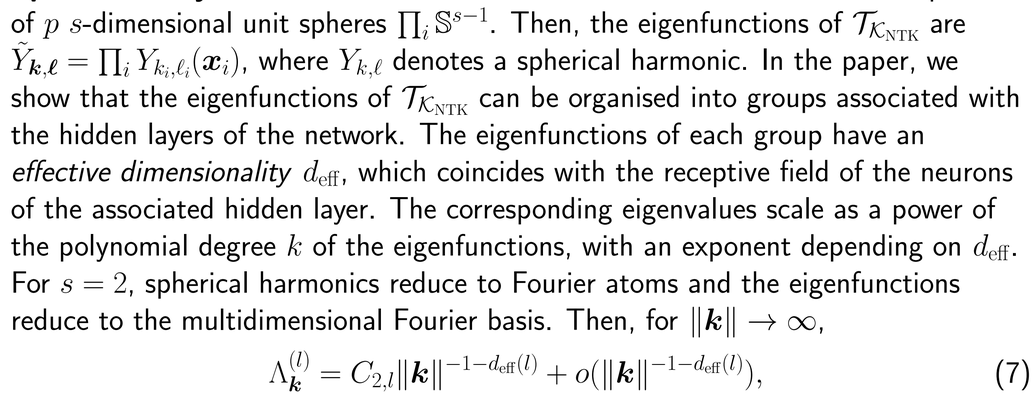

Figure 2. Agreement of theoretical predictions (dashed) with numerical experiments in a teacher (T) student (S) setting where data are generated and learned by CNNs in the kernel limit.

References

[1] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. 2019.

[2] Arthur Jacot, Franck Gabriel, and Clément Hongler. *Neural tangent kernel: Convergence and generalization in neural networks*. 2018.

[3] Sanjeev Arora et al. *On exact computation with an infinitely wide neural net*. 2019.

[4] Andrea Caponnetto and Ernesto De Vito. *Optimal rates for the regularized least-squares algorithm*. 2007.

[5] Francis Bach. *Learning Theory from First Principles*. 2021.

[6] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. *Spectrum dependent learning curves in kernel regression and wide neural networks*. 2020.