Locality defeats the curse of dimensionality in convolutional teacher-student scenarios

Alessandro Favero*, Francesco Cagnetta*, Matthieu Wyart

NEURAL INFORMATION PROCESSING SYSTEMS

EPFL

Problem

- **Supervised learning**
  Approximate a target function $f^*: \mathbb{R}^d \to \mathbb{R}$ from a number P of observations $(x, f^*(x))$ up to a target error ϵ (e.g. mean squared error)

- **How many observations?**
  For a Lipschitz-continuous target $P = \mathcal{O}(\epsilon^{-d})$, i.e.
  $$\epsilon \sim P^{-1/d}$$
  also known as the **curse of dimensionality** (CoD).

- **How many observations in practice?**
  In practice, deep learning algorithms enjoy a much faster decay of the error, e.g. Hastness et al. (2017) report $\epsilon \sim P^{-0.3}$ for ResNets trained on ImageNet ($d = \mathcal{O}(10^5) \Rightarrow 1/d \ll 0.3$)

- **Solution**
  Harness structure of the target function, e.g. **locality**

Image $\boldsymbol{x}$

Patches $\boldsymbol{x}_i$

Local task $f^*(\boldsymbol{x}) = \sum_{i=1}^{d} g_i(\boldsymbol{x}_i)$

Learning scenario

Teacher

Inputs are d-dimensional sequences of uniformly-random numbers $\boldsymbol{x} = (x_1, x_2, x_i, \dots, x_{i+t-1}, \dots, x_d)$ and $\boldsymbol{x}_i = (x_i, x_{i+1}, \dots, x_{i+t-1})$ denotes a t-dimensional patch

The target function is either **local** or **convolutional**

$$f^*_{LOC}(\boldsymbol{x}) = \sum_{i=1}^{d} g_i(\boldsymbol{x}_i), \quad f^*_{CONV}(\boldsymbol{x}) = \sum_{i=1}^{d} g(\boldsymbol{x}_i)$$

Each g_i is a Gaussian random function with controlled smoothness α_t

$$\mathbb{E}[g_i(\boldsymbol{x})] = 0, \quad \mathbb{E}[g_i(\boldsymbol{x}) g_i(\boldsymbol{y})] = C(\boldsymbol{x} - \boldsymbol{y}) \sim \|\boldsymbol{x} - \boldsymbol{y}\|^{\alpha_t}$$

Student

Kernel method with a **local** or **convolutional** kernel with patches of size s and smoothness α_s

$$K^{LOC}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{d} \sum_{i=1}^{d} C(\boldsymbol{x}_i - \boldsymbol{y}_i), \quad K^{CONV}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{d^2} \sum_{i,j=1}^{d} C(\boldsymbol{x}_i - \boldsymbol{y}_j)$$

Predictor from ridge less regression in the corresponding RKHS

$$f = \arg\min_{f \in \mathcal{H}} \sum_{\mu=1}^{P} \left(f(\boldsymbol{x}^\mu) - f^*(\boldsymbol{x}^\mu) \right)^2$$

Generalisation error $\epsilon = \mathbb{E}_{\boldsymbol{x}, f^*} \left(f(\boldsymbol{x}^\mu) - f^*(\boldsymbol{x}^\mu) \right)^2$

Includes infinitely wide nets in the lazy regime!

Spectral bias

Each positive-definite kernel admits Mercer's decomposition

$$K(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\rho > 0} \Lambda_\rho \Phi_\rho(\boldsymbol{x}) \Phi_\rho(\boldsymbol{y})$$

If the target can be expanded in this basis with coefficients c_ρ , and both Λ_ρ and c_ρ decay as inverse powers of ρ (Λ_ρ faster), then from statistical physics [1], [2], [3]

$$\epsilon(P) \sim \sum_{\rho > P} \mathbb{E}[|c_\rho|^2]$$

Similar rates can be obtained rigorously from the spectrum [4] [5]

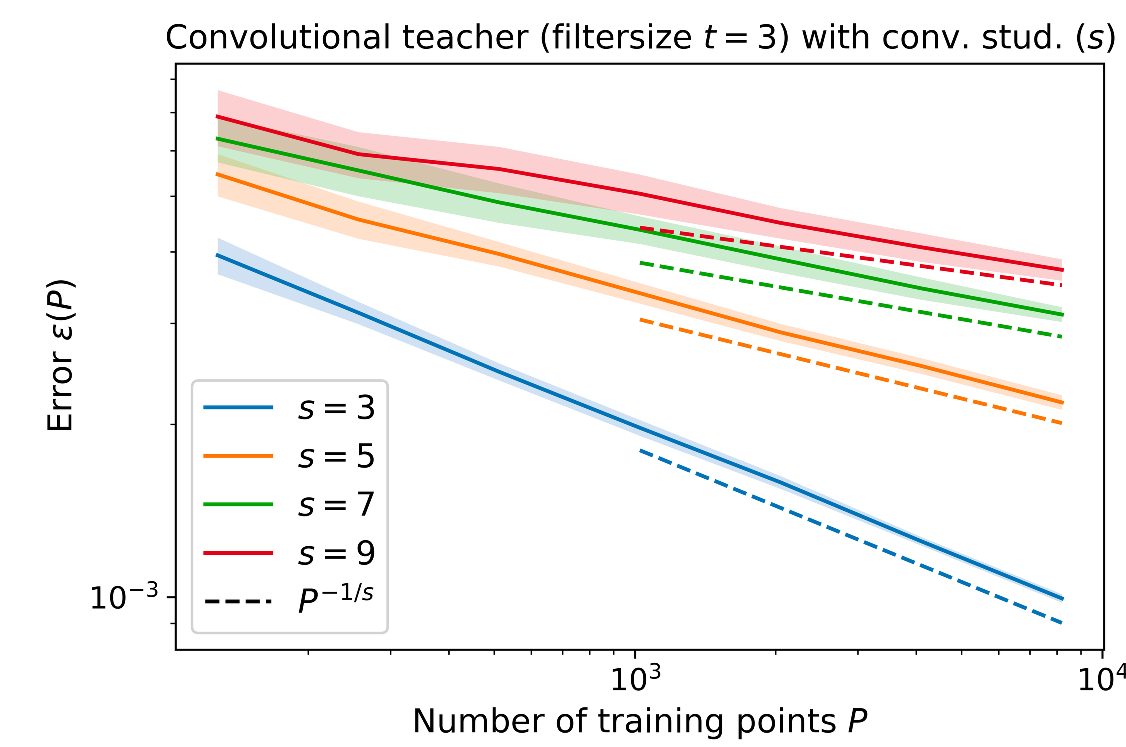
Asymptotic learning curves

Local and convolutional kernels inherit their spectral decomposition form that of the 'smaller' kernel $C(\boldsymbol{x} - \boldsymbol{y})$

$$\Lambda_\rho = \frac{1}{d} \lambda_\rho, \quad \Phi_\rho(\boldsymbol{x}) = \sqrt{\frac{1}{d}} \sum_{i=1}^{d} \phi_\rho(\boldsymbol{x}_i)$$

By the spectral bias, if $s \geq t$ and $\alpha_t \leq 2(\alpha_s + s)$

$$\epsilon(P) \sim P^{-\alpha_t / s}$$



Faster decays than CoD!

Conclusions and perspectives

- Local kernels beat the curse of dimensionality when learning local target functions
- Incorporating shift-invariance in the kernel only yields to pre-asymptotic improvements
- To do: explore the **benefits of depth** studying compositional kernels on hierarchical targets

References

[1] B. Bordelon et al. (2020). "Spectrum dependent ...". In: ICML 2020
[2] A. Canatar et al. (2021). "Spectral bias ...". In: Nat. Comm. 2021
[3] B. Loureiro et al. (2021). "Learning curves ...". In: NeurIPS 2021
[4] A. Jacot et al. (2021). "Kernel alignment...". In: NeurIPS 2020
[5] A. Caponnetto et al. (2006). "Optimal rates...". In: Fou. Com. Mat. 2006

* Equal contribution