How rotational invariance of common kernels prevents generalization in high dimensions

Konstantin Donhauser, Mingqi Wu, Fanny Yang
ETH Zurich – Statistical Machine Learning Group

ICML Conference 2021



Problem setting

- Kernel ridge regression estimate with $\lambda \geq 0$

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

- High dimensional regime where $d, n \to \infty$, $d \asymp n^{\beta}$ with $\beta > 0$

Question: What can we consistently learn in this setting with rotational invariant kernels?

Uniform distributions on spheres

- Ground truth is low degree polynomial

Analysis heavily relies on distribution

General distributions

- Ground truth has bounded Hilbert norm as $d \to \infty$

Unclear for which ground truth this holds

Main results

Theorem *(informal)* – polynomial approximation barrier

$$\mathrm{B}(\hat{f}) \geq \inf_{p \in P_{\leq m}} \mathbb{E}_{X \sim \mathbb{P}_X} (f^\star(X) - p(X))^2 - o(1) \qquad a.s. \text{ as } d, n \to \infty$$

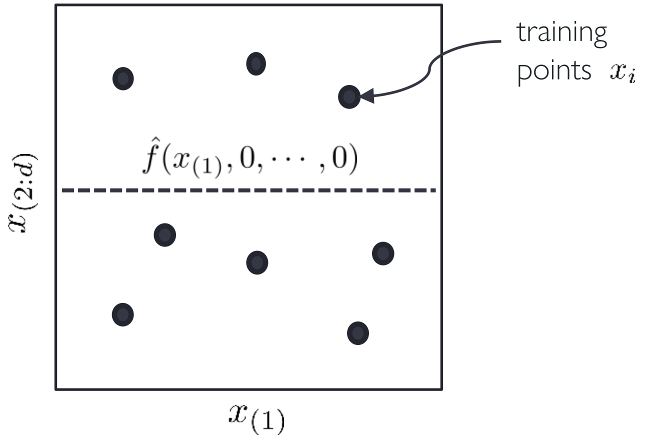
$P_{\leq m}$ set of polynomials of degree at most $2\lfloor 2/\beta \rfloor$

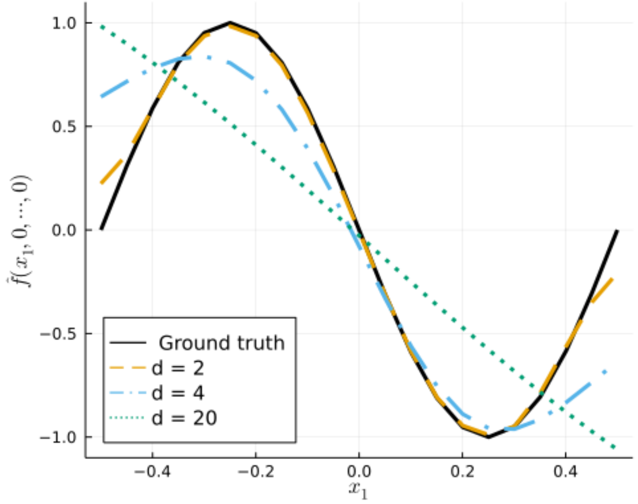
- general input data distribution $\mathbb{P}_X$ with covariance matrix Σ; $d_{\text{eff}} = \mathrm{tr}(\Sigma) \asymp n^\beta$ (for $\Sigma = I_d,\ d \asymp n^\beta$)
- broad range of commonly used rot. inv. kernels of different eigenvalue decays
  - including fully connected NTK of any depth, Laplacian, Gaussian, inner product
- different scalings beyond the classical choice $\tau = d_{\text{eff}}$

Illustration of the polynomial approximation barrier

Experimental details

- $x_i \sim \mathcal{U}([-0.5, 0.5]^d)$
- $y_i = \sin(x_{i,(1)})$
- n =100 iid samples
- Interpolate ($\lambda = 0$)
- Laplace kernel





As dimension grows, the estimator degenerates to a low degree polynomial

Thank you for listening to this talk

We are looking forward to seeing you during the poster session