How rotational invariance of common kernels prevents generalization in high dimensions

Konstantin Donhauser, Mingqi Wu and Fanny Yang
Department of Computer Science, ETH Zurich

Problem Setting

We aim to minimize the population risk of an estimator $\hat{f}$ with $\mathbb{E}_Y$ the expectation over the observation noise during training

$$\mathbf{R}(\hat{f}_\lambda) = \underbrace{\|\mathbb{E}_Y \hat{f}_\lambda - f^\star\|^2_{\mathcal{L}_2(\mathbb{P}_X)}}_{\text{Bias } \mathbf{B}} + \underbrace{\mathbb{E}_Y \|\mathbb{E}_Y \hat{f}_\lambda - \hat{f}_\lambda\|^2_{\mathcal{L}_2(\mathbb{P}_X)}}_{\text{Variance } \mathbf{V}},$$

Given i.i.d. samples $\{x_i, y_i\}_{i=1}^n \sim \mathbb{P}_{X,Y}$, we define the estimators $\hat{f}$

- Kernel ridge regression ($\lambda > 0$)

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

- Kernel interpolation ($\lambda = 0$)

$$\hat{f}_0 = \arg\min_{f \in \mathcal{H}} \|f\|_{\mathcal{H}} \text{ such that } \forall i: \ f(x_i) = y_i$$

High dimensional asymptotics $d, n \to \infty$

- **Covariance model**: We assume that the input data distribution has covariance matrix Σ with effective dimension d_{eff} defined as $d_{\text{eff}} := \text{tr}(\Sigma_d) / \|\Sigma_d\|_{\text{op}}$
- **High dimensional regime**: We assume that the effective dimension grows with the sample size n s.t. $d_{\text{eff}}/n^\beta \to c$ for some $\beta, c > 0$.

Prior Literature

Uniform distributions on spheres [1, 2]

- We can learn polynomials of degree at most $\lfloor 1/\beta \rfloor$ (we call this the polynomial approximation barrier)

General distributions with $\Sigma_d = I_d$ [3]

- Vanishing bias if ground truth function has bounded Hilbert norm as $d \to \infty$
- *Comment:* Unclear when assumption holds

Can we overcome the polynomial approximation barrier when considering different high-dimensional input distributions, eigenvalue decay rates or scalings of the kernel function?

Assumptions

We study rotational invariant kernels of the form

$$k(x.x') = g(\|x\|_2^2, \|x'\|_2^2, x^\top x') = \sum_{j=0}^\infty g_j(\|x\|_2^2, \|x'\|_2^2)(x^\top x')^j$$

- Fully connected NTK of any depth, Laplace kernel, Gaussian kernel, inner product kernels

Scale dependent kernel We scale the data by τ dependent on d, n, i.e. $k_\tau(x, x') = k\left(x/\sqrt{\tau}, x'/\sqrt{\tau}\right)$

- The standard choice $\tau = d_{\text{eff}}$
- Flat limit $\tau \to 0$ (only RBF kernels)

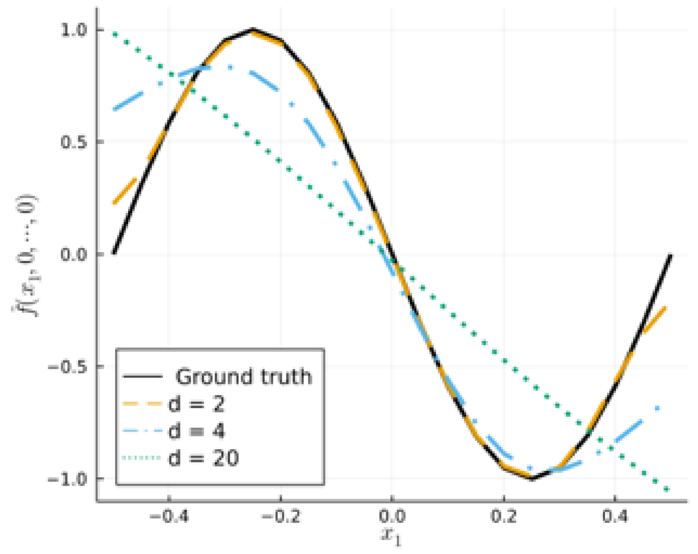
Main Result

Theorem 1. *Polynomial approximation barrier - informal*
Let $\mathcal{P}_{\leq m}$ be the set of polynomials of degree at most $m = 2\lfloor 2/\beta \rfloor$. The bias of the kernel estimators $\hat{f}_\lambda$ with $\lambda \geq 0$ is asymptotically almost surely lower bounded for any $\epsilon > 0$,

$$\mathbf{B}(\hat{f}_\lambda) \geq \inf_{p \in \mathcal{P}_{\leq m}} \|f^\star - p\|_{\mathcal{L}_2(\mathbb{P}_X)} - \epsilon \quad a.s. \text{ as } n \to \infty.$$

Illustration of Polynomial Barrier

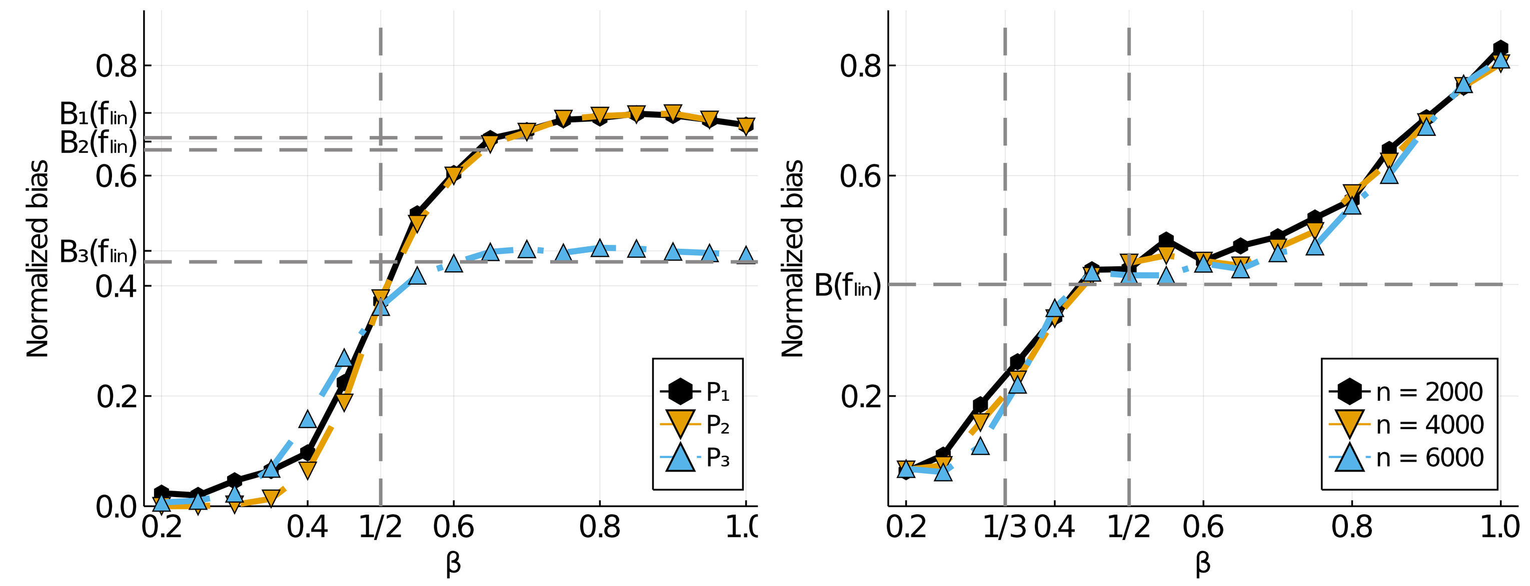
- Interpolate with Laplace kernel
- $n = 100$ i.i.d. samples
- $x_i \sim \text{Uniform}([-0.5, 0.5]^d)$
- $y_i = sin(2\pi x_{i,(1)})$



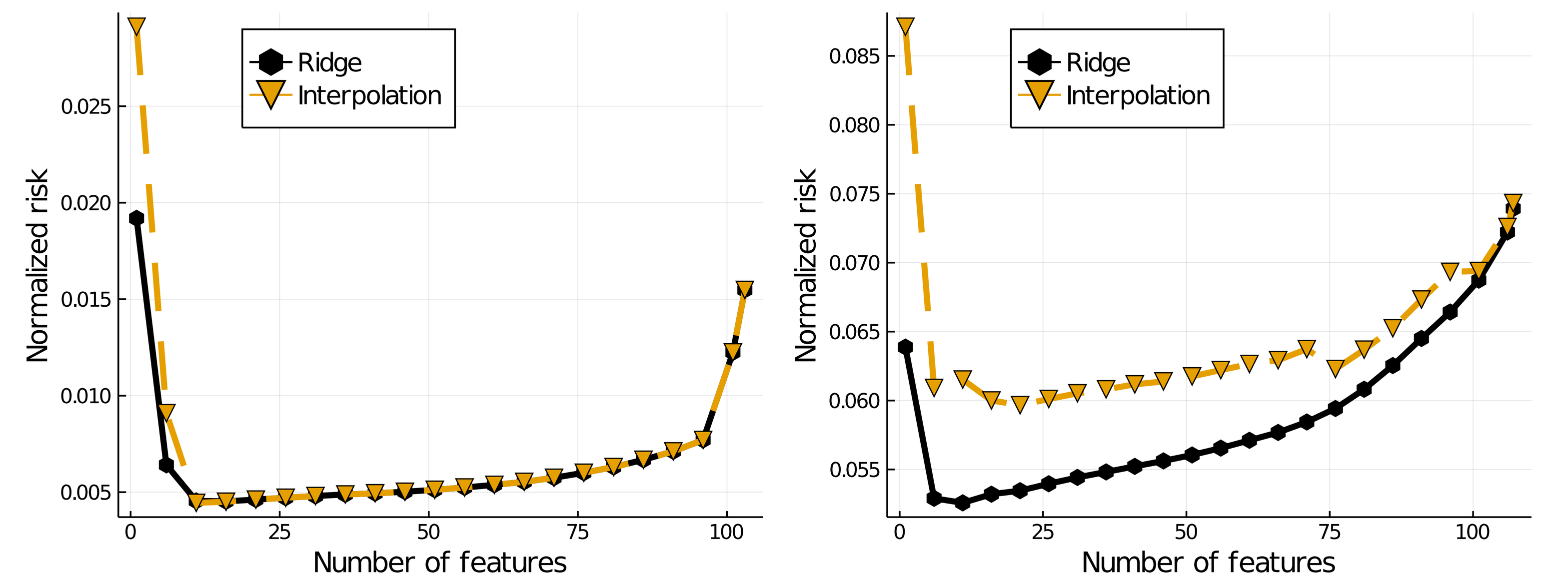
As d increases we observe that $\hat{f}$ degenerates to a linear function

Numerical Results

Synthetic simulations for varying β



Real world dataset (without and with artificial noise)



Discussion and Future Work

Our lower bound applies to

- a broad range of commonly used rotational invariant kernels with **different eigenvalue decays** including exponential (Gaussian kernel) and polynomial (Laplace kernel, NTK)
- input distributions with **general covariance matrices** Σ
- **different scalings** beyond standard choice $\tau = d_{\text{eff}}$

To overcome the polynomial approximation barrier, we therefore propose to investigate in future work how to incorporate prior knowledge to break the rotation invariance

References

[1] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari, "Linearized two-layers neural networks in high dimension," *Annals of Statistics*, vol. 49, no. 2, pp. 1029 – 1054, 2021.
[2] ——, "When do neural networks outperform kernel methods?" in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 14 820–14 830.
[3] T. Liang, A. Rakhlin, and X. Zhai, "On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels," in *Proceedings of the Conference on Learning Theory (COLT)*, 2020.
[4] N. El Karoui *et al.*, "The spectrum of kernel random matrices," *Annals of Statistics*, vol. 38, no. 1, pp. 1–50, 2010.