When Do Neural Networks Outperform Kernel Methods?

Behrooz Ghorbani [1,*] Song Mei [2,*] Theodor Misiakiewicz [3,*] Andrea Montanari [3,4]

[1]Google Research [2]Department of Statistics, UC Berkeley [3]Department of Statistics, Stanford University [4]Department of Electrial Engineering, Stanford University

*Equal contributions

Introduction

For a certain scaling of the initialization (Xavier initialization), sufficiently wide neural networks have been shown to behave like kernel methods, the **Neural Tangent Kernel** [5].

From a theoretical perspective:

- NNs encode a richer class of functions than RKHS.
- Kernel methods can be shown to suffer from the curse of dimensionality

  ... while neural networks can potentially overcome the curse of dimensionality by learning a good low-dimensional representation of the data [1].
- Special examples for which SGD-trained NN provably outperform RKHS methods.

What about in practice? Empirical studies:

- Varied performance gap between the two model classes.
- In some classification tasks, RKHS methods can replace NNs without a large drop in performance.

Can we reconcile these observations?

Focus of this work:

When can we expect a large performance gap between NNs and RKHS methods? For which tasks do NNs outperform RKHS methods?

Spiked Covariates (SC) model

Stylized scenario that captures two properties of datasets:

- Target function depending on a low-dimensional projection;
- Approximately low-dimensional covariates.

Covariates: there exists $[\boldsymbol{U}, \boldsymbol{U}^\perp]$ orthogonal matrix,

$$\boldsymbol{x} = \boldsymbol{U}\boldsymbol{z}_1 + \boldsymbol{U}^\perp \boldsymbol{z}_2.$$

- Signal part: $\boldsymbol{z}_1 \sim \mathsf{Unif}\Big(\mathbb{S}^{d_s-1}\big(\sqrt{\mathsf{snr}_c \cdot d_s}\big)\Big)$.
- Noise part: $\boldsymbol{z}_2 \sim \mathsf{Unif}\Big(\mathbb{S}^{d-d_s-1}\big(\sqrt{d-d_s}\big)\Big)$

$\mathbb{S}^{d-1}(r) = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = r\}$ sphere of radius r in d dimension.

Target function: $f_\star(\boldsymbol{x}) = \varphi(\boldsymbol{z}_1)$.

Parameters of the model:

- Signal dimension: $d_s = d^\eta$, $0 \le \eta \le 1$.
- Covariate SNR: $\mathsf{snr}_c = d^\kappa$, $0 \le \kappa < \infty$ (measures anisotropy of the data, see Fig. 1).

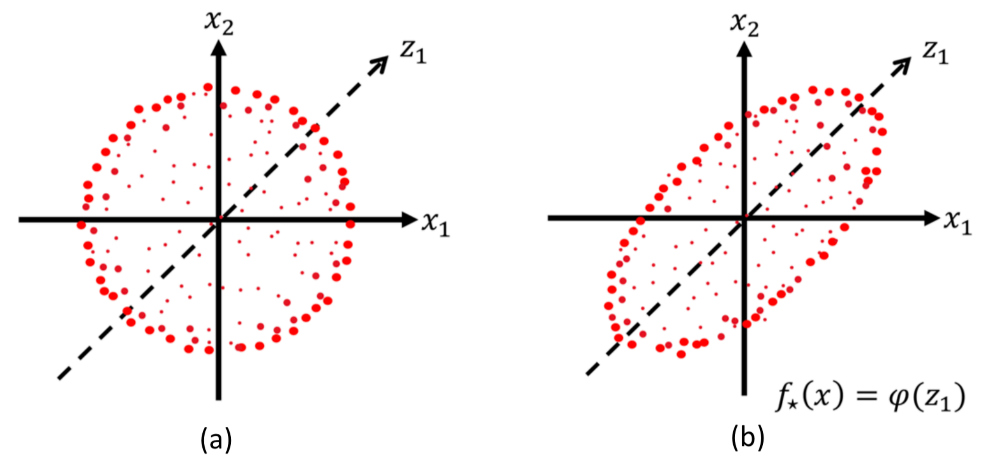

Figure 1: Spiked covariates model: (a) Isotropic covariates ($\kappa = 0$, $\mathsf{snr}_c = 1$). (b) Anisotropic covariates ($\kappa > 0$, $\mathsf{snr}_c > 1$).

Approximation error gap

- Two-layers NNs function class:

$$\mathcal{F}_{\mathsf{NN},N} = \Big\{ f_N(\boldsymbol{x};\boldsymbol{\Theta}) = \sum_{i=1}^N a_i \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle) : a_i \in \mathbb{R}, \boldsymbol{w}_i \in \mathbb{R}^d \Big\}.$$

- Associated neural tangent model: $\mathcal{F}_{\mathsf{RF},N}(\boldsymbol{W}) \oplus \mathcal{F}_{\mathsf{NT},N}(\boldsymbol{W})$ where $\boldsymbol{W} = (\boldsymbol{w}_i)_{i\in[N]} \sim_{iid} \mathsf{Unif}(\mathbb{S}^{d-1})$ are fixed:

$$\mathcal{F}_{\mathsf{RF},N}(\boldsymbol{W}) = \Big\{ f = \sum_{i=1}^N a_i \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle) : a_i \in \mathbb{R}, i \in [N] \Big\},$$

$$\mathcal{F}_{\mathsf{NT},N}(\boldsymbol{W}) = \Big\{ f = \sum_{i=1}^N \langle \boldsymbol{b}_i, \boldsymbol{x}\rangle \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle) : \boldsymbol{b}_i \in \mathbb{R}^d, i \in [N] \Big\}.$$

  Blue: random and fixed. Red: parameters to be optimized.
- With proper initialization, wide NNs trained by GD are well approximated by the neural tangent model [2], [3].

Approximation error for a class of function $\mathcal{F}_N$:

$$R_{\mathsf{App}}(f_\star, \mathcal{F}_N) = \inf_{f \in \mathcal{F}_N} \mathbb{E}_{\boldsymbol{x}}\Big[\Big(f_\star(\boldsymbol{x}) - f(\boldsymbol{x})\Big)^2\Big].$$

Effective dimension: $d_{\mathsf{eff}} = d_s \vee (d/\mathsf{snr}_c)$.

Approximation error in SC model

Theorem 1 ([4]) *Assume $d_{\mathsf{eff}}{}^{\ell+\delta} \le N \le d_{\mathsf{eff}}{}^{\ell+1-\delta}$ and σ satisfies "generic conditions". Then*

$$R_{\mathsf{App}}(f_\star, \mathcal{F}_{\mathsf{RF},N}(\boldsymbol{W})) = \|\mathsf{P}_{>\ell} f_\star\|_{L^2}^2 + o_{d,\mathbb{P}}(\cdot),$$
$$R_{\mathsf{App}}(f_\star, \mathcal{F}_{\mathsf{NT},N}(\boldsymbol{W})) = \|\mathsf{P}_{>\ell+1} f_\star\|_{L^2}^2 + o_{d,\mathbb{P}}(\cdot).$$

On the contrary, assume $d_s{}^{\ell+\delta} \le N \le d_s{}^{\ell+1-\delta}$, we have

$$R_{\mathsf{App}}(f_\star, \mathcal{F}_{\mathsf{NN},N}) \le \|\mathsf{P}_{>\ell+1} f_\star\|_{L^2}^2 + o_d(\cdot).$$

Furthermore, $R_{\mathsf{App}}(f_\star, \mathcal{F}_{\mathsf{NN},N})$ is independent of snr_c.

$\mathsf{P}_{>\ell}$*: projection orthogonal to the space of degree-ℓ polynomials.*

- d_{eff}: capture the "effective low-dimensionality" of the data.
- For RF/NT, random $\boldsymbol{w}_i$'s have small correlation with $\boldsymbol{z}_1$ in high dimension. This is alleviated by higher snr_c.
- For NN, $\boldsymbol{w}_i$'s can be chosen with large correlation with $\boldsymbol{z}_1$.
- NN can "adaptively learn" $\boldsymbol{w}_i$'s while RF/NT cannot.

Generalization error gap

- Kernel Ridge Regression: given a rotationally invariant kernel $H(\boldsymbol{x}, \boldsymbol{y}) = h(\langle \boldsymbol{x}, \boldsymbol{y}\rangle)$ and regularization λ,

$$\hat{\boldsymbol{a}}^\lambda := \arg\min_{\boldsymbol{a}\in\mathbb{R}^n} \Big\{ \frac{1}{n}\sum_{i=1}^n \Big(y_i - \sum_{i=1}^n a_i h(\langle \boldsymbol{x}, \boldsymbol{x}_i\rangle)\Big)^2 + \lambda \boldsymbol{a}^\mathsf{T} \boldsymbol{H} \boldsymbol{a} \Big\}.$$

  and the solution $\hat{f}_{h,n,\lambda}(\boldsymbol{x}) = \sum_{i=1}^n \hat{a}_i^\lambda h(\langle \boldsymbol{x}, \boldsymbol{x}_i\rangle)$.
- NTK with any number of layers with iid Gaussian initialization is rotationally invariant.
- Generalization error:

$$R_{\mathsf{Gen}}(f_\star, \hat{f}_{h,n,\lambda}) = \mathbb{E}_{\boldsymbol{x}}\Big[\Big(f_\star(\boldsymbol{x}) - \sum_{i=1}^n \hat{a}_i^\lambda h(\langle \boldsymbol{x}, \boldsymbol{x}_i\rangle)\Big)^2\Big]$$

Generalization error in SC model

Theorem 2 ([4]) *Assume $d_{\mathsf{eff}}{}^{\ell+\delta} \le n \le d_{\mathsf{eff}}{}^{\ell+1-\delta}$, $h(\cdot)$ satisfies "generic conditions" and $\lambda = O_d(1)$. Then*

$$R_{\mathsf{Gen}}(f_\star, \hat{f}_{h,n,\lambda}) = \|\mathsf{P}_{>\ell} f_\star\|_{L^2}^2 + o_{d,\mathbb{P}}(\cdot).$$

$\mathsf{P}_{>\ell}$*: projection orthogonal to the space of degree-ℓ polynomials.*

- What about NNs trained by GD? Currently out of reach.
- We can construct a NN (PCA on $(\boldsymbol{x}_i)_{i\in[n]}$ + training on the subsphere) such that for $d_s{}^{\ell+\delta} \le n \le d_s{}^{\ell+1-\delta}$,

$$R_{\mathsf{Gen}}(f_\star, \hat{f}_{NN,N}) = \|\mathsf{P}_{>\ell} f_\star\|_{L^2}^2 + o_{d,\mathbb{P}}(\cdot).$$

- In some cases, we expect the performance of NNs trained in the mean-field regime to depend on d_s and not d (empirical and theoretical evidence supporting this conjecture).

Summary

We have d_{eff} decreases with snr_c:

- Small snr_c ($d_{\mathsf{eff}} = d$): isotropic covariates,

  Approximation error: $\mathsf{NN} \ll \mathsf{RF/NT}$,

  Generalization error: $\mathsf{NN} \ll \mathsf{KRR}$.
- Large snr_c ($d_{\mathsf{eff}} = d_s$): highly anisotropic covariates,

  Approximation error: $\mathsf{NN} \sim \mathsf{RF/NT}$,

  Generalization error: $\mathsf{NN} \sim \mathsf{KRR}$.

In this stylized model, a controlling parameter of the performance gap between NN and kernel methods is

$$\mathsf{snr}_c = \frac{\text{Signal covariates variance}}{\text{Noise covariates variance}}.$$

Latent low-dimensional structure in the covariates and the target function alleviates the curse of dimensionality and make kernel methods more competitive.

Testing insights on real datasets

In *image classification*, we expect

- The labels to depend predominantly on the low-frequency components of the images;
- Spectrum of images to concentrate on low-frequencies.

Insight I: lower covariate SNR (data more isotropic) should lead to larger generalization gap between NN and RKHS.

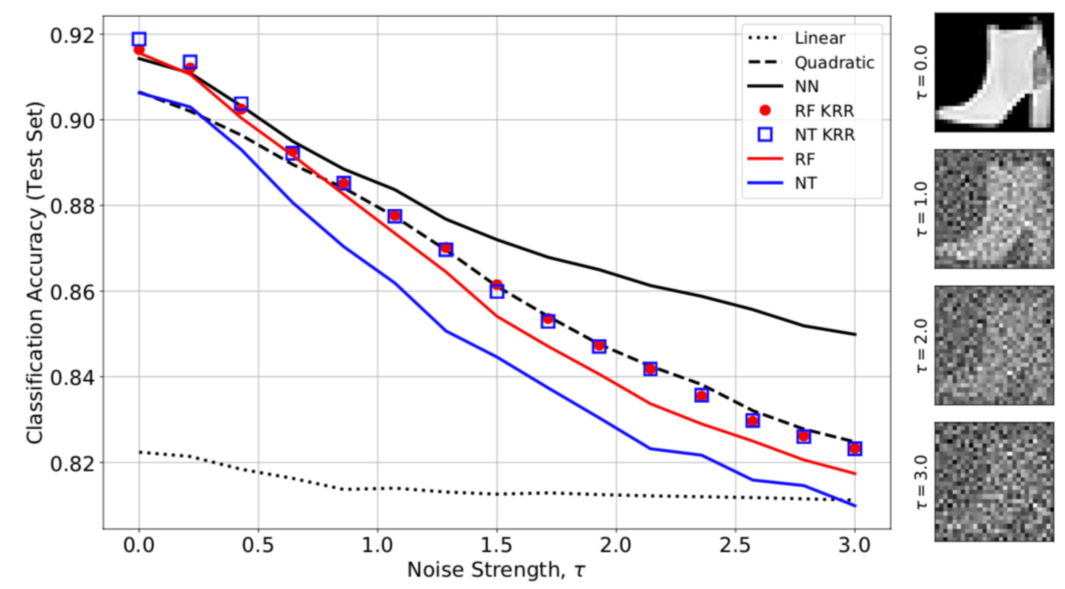

Figure 2: Test accuracy on Fashion MNIST: adding noise to the high frequency components (decreases snr_c).

Insight II: if low-dimensional structure of the target function is not aligned with low-dimensional covariates, we should expect a larger generalization gap between NN and RKHS.

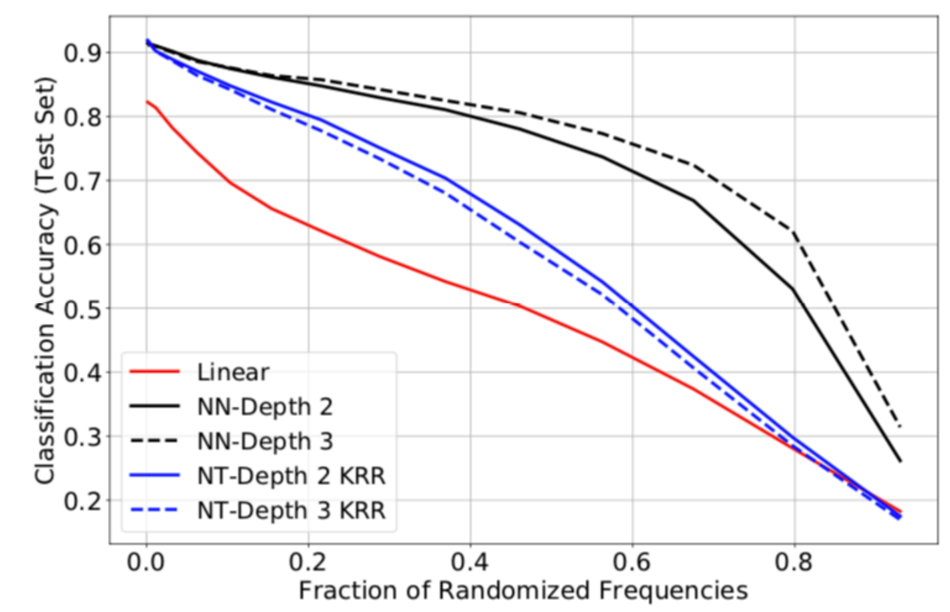

Figure 3: Test accuracy on Fashion MNIST: replacing the low-frequency components by noise with matching covariance (de-align the labels from the low-frequency components).

Bibliography

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

[2] L. Chizat, E. Oyallon, and F. Bach.
On lazy training in differentiable programming.
In *Advances in Neural Information Processing Systems*, pages 2937–2947, 2019.

[3] S. S. Du, J. D. Lee, H. Li, L. Wang, and X. Zhai.
Gradient descent finds global minima of deep neural networks.
arXiv:1811.03804, 2018.

[4] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari.
When do neural networks outperform kernel methods?
In *Advances in Neural Information Processing Systems*, 2020.

[5] A. Jacot, F. Gabriel, and C. Hongler.
Neural tangent kernel: Convergence and generalization in neural networks.
In *Advances in neural information processing systems*, pages 8571–8580, 2018.