A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent

NEURAL INFORMATION PROCESSING SYSTEMS

Zhenyu Liao (1), Romain Couillet (2), Michael W. Mahoney (1)

(1) ICSI and Department of Statistics, University of California, Berkeley, USA; (2) G-STATS Data Science Chair, GIPSA-lab, University Grenoble-Alpes, France.

Abstract

Questions:

- Random Fourier features (RFFs) approximate Gaussian kernel, but in which sense? Entry-wise vs. operator norm?
- Large dimensional machine learning systems establish *double descent* test curves, is this true for **any** data?

Results:

- RFF Gram matrix approximates Gaussian kernel *entry-wise*, and in *operator norm* **only** when the number N of features $\gg$ number n of samples.
- **Sharp** analysis of RFF regression performance for **any** N/n.
- Double descent proved to exist on **real-world** data!

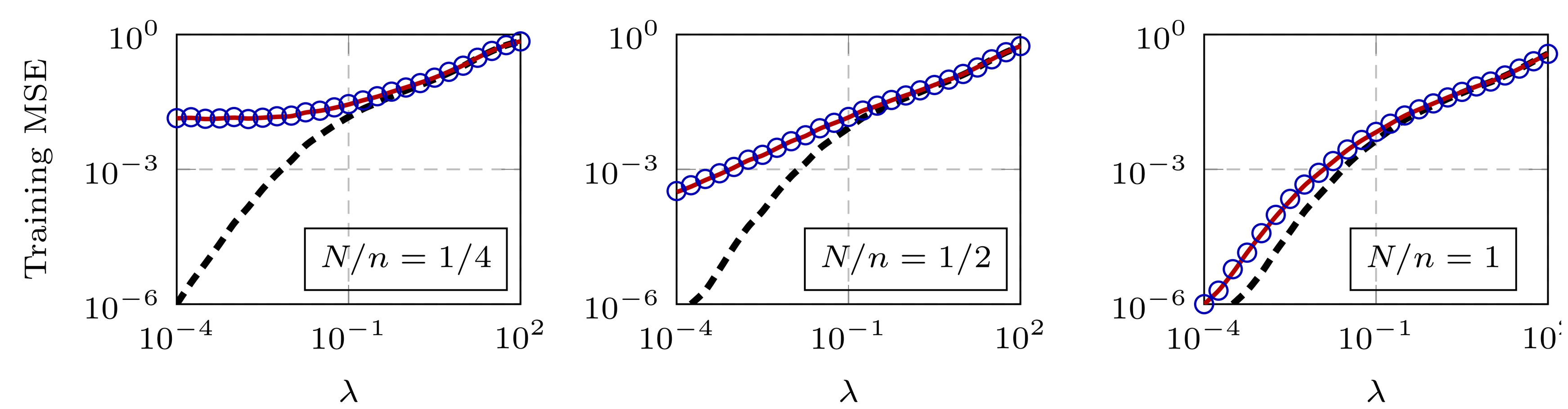
Entry-wise ≠ operator norm convergence

Setup:

- RFF $\Sigma = \sigma(WX) \in R^{N \times n}$ for n data points $X = [x_1, \dots, x_n] \in R^{p \times n}$ of dimension p, with random (e.g., Gaussian) matrix $W \in R^{N \times p}$.
- [1]: *entry-wise* convergence of RFF Gram matrix $\frac{1}{N}[\Sigma^T \Sigma]_{ij} \to [K_{Gauss}]_{ij}$ Gaussian kernel as $N \to \infty$.
- **Not true** for *operator norm* $\|\frac{1}{N}\Sigma^T\Sigma - K_{Gauss}\| \gg 0$ unless $N \gg n$.
- **Example**: for $N < n$, $\Sigma^T\Sigma \in R^{n \times n}$ of rank **at most** N and has **at least** $n - N$ zero eigenvalues, while K_{Gauss} of *full rank* for distinct $x_1, \dots, x_n \Rightarrow$ eigenvalue **mismatch!**
- Due to $\|A\|_\infty \le \|A\| \le n\|A\|_\infty$, $\|A\|_\infty = \max_{ij} |A|_{ij}$.
- Significant impact on various RFF-based algorithms.

ICSI Berkeley UNIVERSITY OF CALIFORNIA UNIVERSITÉ Grenoble Alpes

Sharp analysis of RFF regression via RMT



Training MSEs of RFF ridge regression on MNIST data (class 3 versus 7) as a function of regression penalty λ.

- Theoretical guarantee from Gauss kernel (**black** dashed lines) **different** from empirical observations (**blue** circles) when N is not much larger than n.
- Random matrix theory (RMT) predictions (**red** lines) **consistently** match empirical results, based on RFF **resolvent**.
- Training set (X, y) RFF resolvent $Q(\lambda) = \left(\frac{1}{n}\Sigma^T\Sigma + \lambda I_n\right)^{-1}$ for $\Sigma_X^T = [\cos(WX)^T, \sin(WX)^T] \in R^{n \times 2N}$, RFF ridge regressor $\beta = \frac{1}{n}\Sigma_X Q(\lambda) y \in R^{2N}$ from minimizing $\frac{1}{n}\|y - \Sigma_X^T\beta\|^2 + \lambda\|\beta\|^2$.

Theorem: Asymptotic equivalent for $E[Q(\lambda)]$

As $n, p, N \to \infty$ at the same pace with $\|X\| = O(1)$ and $\|y\|_\infty = O(1)$,

$$\|E[Q] - \bar{Q}\| \to 0, \quad \bar{Q} = \left(\frac{N}{n}\frac{K_{\cos}}{1+\delta_{\cos}} + \frac{N}{n}\frac{K_{\cos}}{1+\delta_{\sin}} + \lambda I_n\right)^{-1}$$

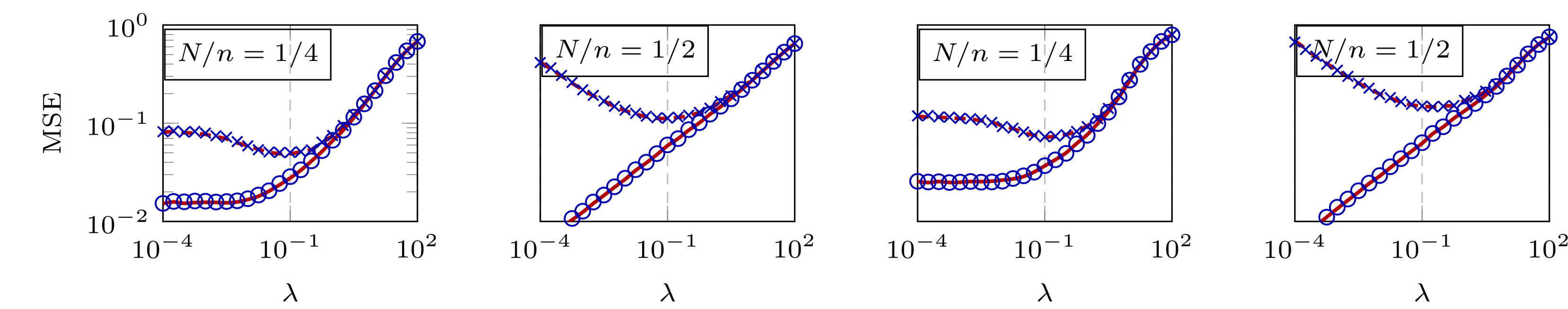
with $[\mathrm{K}_{\cos}]_{ij} = \exp\left(-\frac{\|x_i\|^2 + \|x_j\|^2}{2}\right)\cosh(x_i^T x_j)$ and $[\mathrm{K}_{\sin}]_{ij} = \exp\left(-\frac{\|x_i\|^2 + \|x_j\|^2}{2}\right)\sinh(x_i^T x_j)$ such that $\mathrm{K}_{\cos} + \mathrm{K}_{\sin} = K_{Gauss}$, for $(\delta_{\cos}, \delta_{\sin})$ unique positive solution to

$$\delta_{\cos} = \frac{1}{n} tr\,(K_{\cos}\,\bar{\mathrm{Q}}), \qquad \delta_{\sin} = \frac{1}{n} tr\,(K_{\sin}\,\bar{\mathrm{Q}}).$$

- RFF *Effective kernel* as weighted sum of $K_{\cos}$ and $K_{\sin}$.
- Access to **training** and **test** mean squared errors (MSEs).
- Data dependent theory with **no** strong assumption on the data.

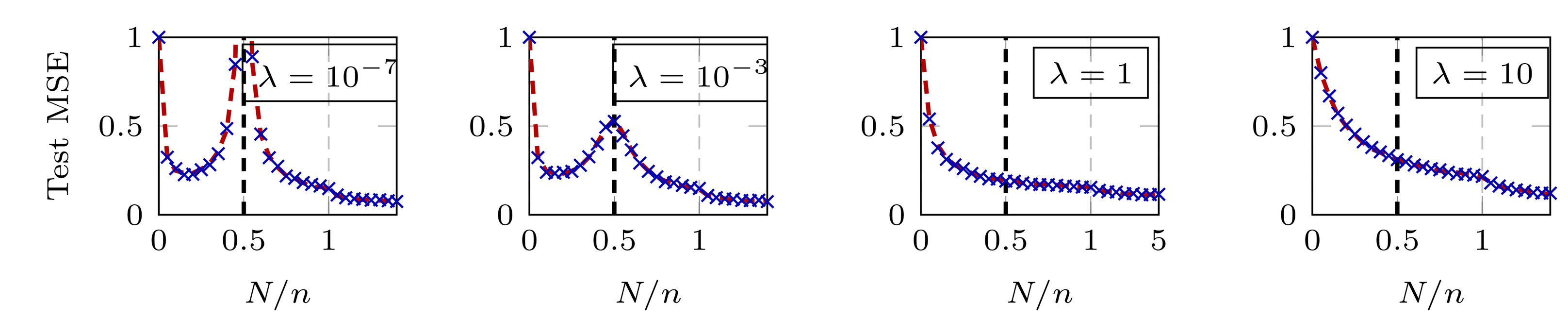
Practical consequences

- Theoretical performance guarantees for RFF ridge regression for **any** ratio N/n on **real-world** data.



MSEs of RFF ridge regression on Fashion- (**left two**) and Kannada-MNIST (**right two**)

- Double descent on real-world data? Yes, **proved** here for RFF!
- Due to an under- to over-parameterization *phase transition* of $\bar{Q}$ in the ridgeless $\lambda \to 0$ limit.



Test MSEs of RFF regression as a function of the ratio N/n, on MNIST data set

Future work

- Property of *effective kernel*? Positive-definite: $\frac{K_{\cos}}{1+\delta_{\cos}} + \frac{K_{\cos}}{1+\delta_{\sin}} \succcurlyeq \frac{K_{Gauss}}{1+\max(\delta_{\cos}, \delta_{\sin})}$. Eigenvalue decay and eigen-structure?
- Better design of random feature based methods by wisely combining different nonlinear activations [4,5].

References

[1] Rahimi, Ali, and Benjamin Recht. "Random features for large-scale kernel machines." In Advances in neural information processing systems, pp. 1177-1184. 2008.

[2] Liao, Zhenyu, Romain Couillet, and Michael W. Mahoney. "**A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent.**" arXiv preprint arXiv:2006.05013 (2020).

[3] Louart, Cosme, Zhenyu Liao, and Romain Couillet. "A random matrix approach to neural networks." The Annals of Applied Probability 28, no. 2 (2018): 1190-1248.

[4] Mei, Song, and Andrea Montanari. "The generalization error of random features regression: Precise asymptotics and double descent curve." arXiv preprint arXiv:1908.05355 (2019).

[5] Liao, Zhenyu, and Romain Couillet. "On the Spectrum of Random Features Maps of High Dimensional Data." In International Conference on Machine Learning, pp. 3063-3071. 2018.