A Data-dependent Theory of Overparameterization: Phase Transition, Double Descent, and Beyond

Workshop on the Theory of Overparameterized Machine Learning (TOPML)

Zhenyu Liao

joint work with Romain Couillet@U Grenoble-Alpes and Michael Mahoney@UC Berkeley

School of Electronic Information and Communications, Huazhong University of Science & Technology

April 21, 2021

Sample covariance matrix in the large n, p regime

- For $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, estimate population covariance $\mathbf{C} \in \mathbb{R}^{p \times p}$ from n data samples $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$.

- Maximum likelihood sample covariance matrix with entry-wise convergence

$$\hat{\mathbf{C}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}} = \frac{1}{n} \mathbf{X}\mathbf{X}^{\mathsf{T}} \in \mathbb{R}^{p \times p}, \quad [\hat{\mathbf{C}}]_{ij} \to [\mathbf{C}]_{ij}$$

  almost surely as $n \to \infty$ (law of large numbers): optimal for $n \gg p$ (or, for p "small").

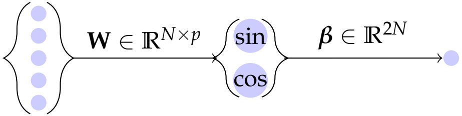
- In the $n \sim p$ regime, conventional wisdom breaks down: for $\mathbf{C} = \mathbf{I}_p$ with $n < p$, $\hat{\mathbf{C}}$ has at least $p - n$ zero eigenvalues.

$$\boxed{\|\hat{\mathbf{C}} - \mathbf{C}\| \not\to 0, \quad n, p \to \infty}$$

  $\Rightarrow$ eigenvalue mismatch and NOT consistent!

- due to $\|\mathbf{A}\|_\infty \leq \|\mathbf{A}\| \leq p\|\mathbf{A}\|_\infty$ for $\mathbf{A} \in \mathbb{R}^{p \times p}$ and $\|\mathbf{A}\|_\infty \equiv \max_{ij} |\mathbf{A}_{ij}|$.
- Marčenko-Pastur law and many fundamental results in **Random Matrix Theory**!

Remainder on random Fourier features (RFFs)



$\mathbf{X} \in \mathbb{R}^{p \times n}$ $\quad \boldsymbol{\Sigma}^\mathsf{T} = [\cos(\mathbf{W}\mathbf{X})^\mathsf{T}, \ \sin(\mathbf{W}\mathbf{X})^\mathsf{T}]$

Figure: Illustration of RFFs regression model.

- random Fourier features $\boldsymbol{\Sigma} \in \mathbb{R}^{2N \times n}$ of data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ with standard Gaussian $\mathbf{W} \in \mathbb{R}^{N \times p}$, i.e., $\mathbf{W}_{ij} \sim \mathcal{N}(0,1)$
- RFF ridge regressor $\boldsymbol{\beta} \in \mathbb{R}^{2N}$ given by

$$\boldsymbol{\beta} \equiv \frac{1}{n}\boldsymbol{\Sigma}\left(\frac{1}{n}\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma} + \lambda \mathbf{I}_n\right)^{-1}\mathbf{y} \cdot 1_{2N>n} + \left(\frac{1}{n}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\mathsf{T} + \lambda \mathbf{I}_{2N}\right)^{-1}\frac{1}{n}\boldsymbol{\Sigma}\,\mathbf{y} \cdot 1_{2N<n}$$

- equivalent for any $\lambda > 0$
- a stylized model for the analysis of neural nets

Random Fourier features imply Gaussian kernel, but in which sense?

- [RR08]: entry-wise convergence of RFF Gram $\frac{1}{N}[\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma}]_{ij} \to [\mathbf{K}_{\text{Gauss}}]_{ij}$ Gaussian kernel matrix as $N \to \infty$
  - Proof: (again) law of large numbers, for $\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$,

$$\frac{1}{N}[\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma}]_{ij} = \frac{1}{N}\sum_{t=1}^{N}\left[\cos(\mathbf{x}_i^\mathsf{T}\mathbf{w}_t)\cos(\mathbf{w}_t^\mathsf{T}\mathbf{x}_j) + \sin(\mathbf{x}_i^\mathsf{T}\mathbf{w}_t)\sin(\mathbf{w}_t^\mathsf{T}\mathbf{x}_j)\right]$$

$$\to \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_p)}\left[\cos(\mathbf{x}_i^\mathsf{T}\mathbf{w})\cos(\mathbf{w}^\mathsf{T}\mathbf{x}_j) + \sin(\mathbf{x}_i^\mathsf{T}\mathbf{w})\sin(\mathbf{w}^\mathsf{T}\mathbf{x}_j)\right] = [\mathbf{K}_{\text{Gauss}}]_{ij}$$

- similar to sample covariance: **not true** in spectral norm, $\|\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma}/N - \mathbf{K}_{\text{Gauss}}\| \not\to 0$ unless $2N \gg n$
  - e.g., $\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ of rank at most $2N$ if $2N \le n$, while $\mathbf{K}_{\text{Gauss}}$ of rank n (for distinct $\mathbf{x}_i$)
  - significant impact on various RFF-based algorithms

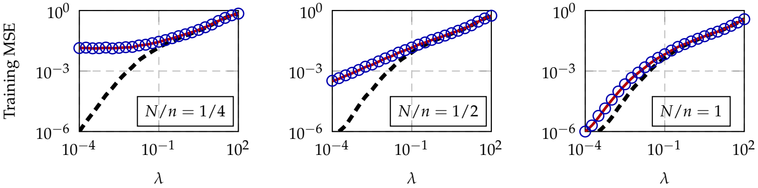


Figure: Training MSEs of RFF ridge regression on MNIST data (class 3 versus 7) as a function of regression penalty λ.

[1] Ali Rahimi and Benjamin Recht. "Random features for large-scale kernel machines". In: *Advances in neural information processing systems.*

Sharp analysis of RFF ridge regression performance via RMT

- provides precise training and test performances of RFF for any ratio N/n and (almost) real data

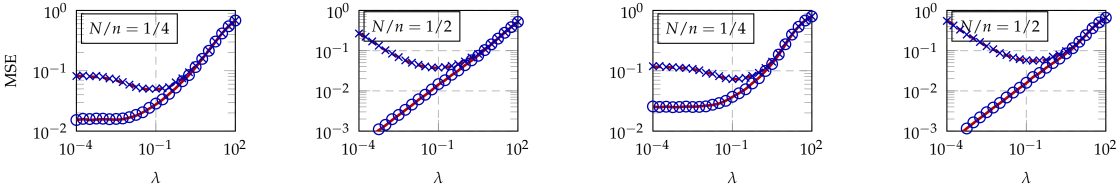


Figure: MSEs of RFF ridge regression on Fashion- (**left two**) and Kannada-MNIST (**right two**).

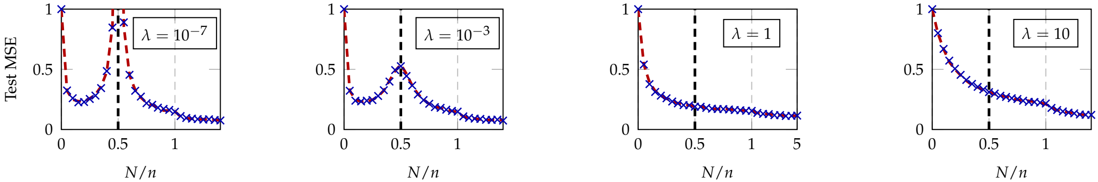


Figure: Test MSEs of RFF regression as a function of the ratio N/n, on MNIST data set.

Conclusion and take-away message

- double descent test curves on real-world data? Yes, **proved** here for RFF!
- phase transition from under- to over-param of resolvent $(\frac{1}{n}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma} + \lambda \mathbf{I}_n)^{-1}$ in the ridgeless $\lambda \to 0$ limit

Take-away message:

- entry-wise $\neq$ spectral norm convergence for large matrices:
  - inconsistency of sample covariance matrix in high dimensions $p \sim n$
  - random Fourier feature maps $\neq$ Gaussian kernel if $N \not\gg n$
- **RMT** provides precise prediction of overparameterized ML algorithms on real-world data!

Ref: "A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent" (NeurIPS 2020, `https://arxiv.org/abs/2006.05013`) and my homepage `https://zhenyu-liao.github.io/` for more information!

Thank you!