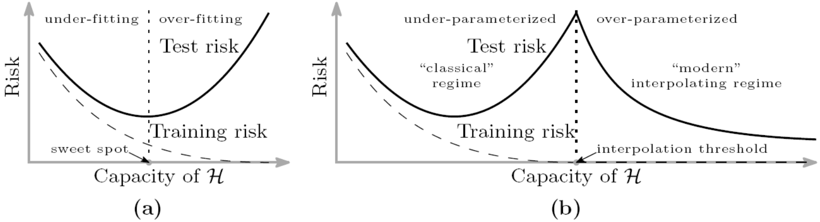
A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent

(Journal of Statistical Mechanics: Theory and Experiment)

Liao et al.
presented by Jihu Lee
January 18, 2022

Table of Contents

Double Descent (Belkin et al. 2019)



- The double descent risk curve can be observed in various situations
- Can be discussed in the context of neural networks: RFF,. . .

Random Fourier Features (Rahimi, Recht, et al. 2007)

- An approach to scaling up kernel methods for shift-invariant kernels
- $E[z_\omega(\mathrm{x})^T z_\omega(\mathrm{y})] = k(\mathrm{x}, \mathrm{y})$

Algorithm 1: Random Fourier Features

Require: A positive definite shift-invariant kernel $k(\mathrm{x}, \mathrm{y}) = k(\mathrm{x} - \mathrm{y})$

Ensure: A randomized feature map $\mathrm{z}(\mathrm{x}) : \mathbb{R}^d \to \mathbb{R}^{2D}$ so that $\mathrm{z}(\mathrm{x})^T \mathrm{z}(\mathrm{y}) \sim k(\mathrm{x} - \mathrm{y})$

Compute the Fourier transform p of the kernel k:

$p(\omega) = \frac{1}{2\pi} \int e^{-i\omega^T \Delta} k(\Delta) d\Delta$

Draw D iid samples $\omega_1, \ldots, \omega_D \in \mathbb{R}^d$ from p

Let $\mathrm{z}(\mathrm{x}) = \sqrt{\frac{1}{D}}[\cos(\omega_1^T \mathrm{x}) \ldots \cos(\omega_D^T \mathrm{x}) \sin(\omega_1^T \mathrm{x}) \ldots \sin(\omega_D^T \mathrm{x}))]^T$

Table of Contents

Introduction

- Asymptotic analysis in classical learning theory:
  $n \to \infty$ for fixed N / $N \to \infty$ for fixed n (less practical!)
- Double asymptotic regime: $n, N \to \infty$ with $N/n \to c$, a constant $\to$ deal with the relative complexity and gives a precise description of the under- to over-parameterized phase transition

Basic Settings

- Considers random feature maps: may be viewed also as a 2-layer neural network
- $\mathrm{X} = [\mathrm{x}_1, \ldots, \mathrm{x}_n] \in \mathbb{R}^{p \times n}$: data matrix
- $\Sigma_{\mathrm{x}} = \sigma(\mathrm{WX}) \in \mathbb{R}^{N \times n}$: random feature map
- $\Sigma_{\mathrm{x}}^T = [\cos(\mathrm{WX})^T,\ \sin(\mathrm{WX})^T)] \in \mathbb{R}^{n \times 2N}$: RFF

Main Contributions

- Precise characterization of the asymptotics of the RFF empirical Gram matrix when n, p, N goes to the limit
- Derive the asymptotic training/test MSE of RFF ridge regression as a function of the ratio N/n and λ
- Detailed empirical evaluation of theoretical results

Table of Contents

Main Technical Results

Assumption

As $n \to \infty$, we have

1. $0 < \liminf_n \min\left\{\frac{p}{n}, \frac{N}{n}\right\} \leq \limsup_n \max\left\{\frac{p}{n}, \frac{N}{n}\right\} < \infty$
2. $\limsup_n \|X\| < \infty$ and $\limsup_n \|y\|_\infty < \infty$

- $\Sigma_X^T = [cos(WX)^T, sin(WX)^T]$ with $W_{ij} \sim \mathcal{N}(0,1)$
- $E_{\text{train}} = \frac{1}{n}\|y - \Sigma_X^T \beta\|^2$, $E_{\text{test}} = \frac{1}{\hat{n}}\|\hat{y} - \Sigma_{\hat{X}}^T \beta\|^2$

Main Technical Results

- RFF ridge regressor is given by,

$$\beta = \frac{1}{n}\Sigma_X(\frac{1}{n}\Sigma_X^T\Sigma_X + \lambda I_n)^{-1}y \cdot 1_{2N>n} + (\frac{1}{n}\Sigma_X\Sigma_X^T + \lambda I_{2N})^{-1}\frac{1}{n}\Sigma_X y \cdot 1_{2N<n}$$

- Resolvent: $Q(\lambda) = (\frac{1}{n}\Sigma_X^T\Sigma_X + \lambda I_n)^{-1}$
- $\frac{1}{N}[\Sigma_X^T\Sigma_X]_{ij} \xrightarrow{a.s.} [K_X]_{ij}$ where K_X is the Gaussian kernel matrix $\left\{exp(-\|x_i - x_j\|^2/2)\right\}_{i,j=1}^n$
- But the convergence in spectral norm $\|\Sigma_X^T\Sigma_X/N - K_X\| \to 0$ does not hold

Asymptotic Deterministic Equivalent

- Need to find a deterministic equivalent for $\mathsf{Q}(\lambda)$
- $\mathbb{E}_{\mathsf{W}}[\mathsf{Q}]$: matrix inverse, not convenient
- Alternative: close to $\mathbb{E}_{\mathsf{W}}[\mathsf{Q}]$ when $n, p, N \to \infty$ and numerically more accessible

Asymptotic Deterministic Equivalent

Theorem (Asymptotic equivalent for $\mathbb{E}_W[Q]$)

Under Assumption 1, for $\lambda > 0$, we have, as $n \to \infty$

$$\|\mathbb{E}_{\mathbf{W}}[Q] - \bar{Q}\| \to 0$$

for $\bar{Q} \equiv (\frac{N}{n}(\frac{K_{\boldsymbol{cos}}}{1+\delta_{\boldsymbol{cos}}} + \frac{K_{\boldsymbol{sin}}}{1+\delta_{\boldsymbol{sin}}}) + \lambda I_n)^{-1}$, $K_{\boldsymbol{cos}} \equiv K_{\boldsymbol{cos}}(X, X)$, $K_{\boldsymbol{sin}} \equiv K_{\boldsymbol{sin}}(X, X) \in \mathbb{R}^{n \times n}$ *and*

$$K_{\boldsymbol{cos}}(X, X')_{ij} = e^{-\frac{\|x_i\|^2 + \|x_j'\|^2}{2}} \cosh(x_i^T x_j'), K_{\boldsymbol{sin}}(X, X')_{ij} = e^{-\frac{\|x_i\|^2 + \|x_j'\|^2}{2}} \sinh(x_i^T x_j')$$

where $(\delta_{\boldsymbol{cos}}, \delta_{\boldsymbol{sin}})$ *is the unique positive solution to*

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

Asymptotic Training Performance

Theorem (Asymptotic Training Performance)

Under Assumption 1, for a given training set (X, y) *and training MSE, as* $n \to \infty$

$$E_{train} - \bar{E}_{train} \xrightarrow{a.s.} 0$$

$$\bar{E}_{train} = \frac{\lambda^2}{n}\|\bar{\mathrm{Q}}\mathrm{y}\|^2 + \frac{N}{n}\frac{\lambda^2}{n^2}\begin{bmatrix}\frac{tr(\bar{\mathrm{Q}}\mathrm{K}_{\boldsymbol{cos}}\bar{\mathrm{Q}})}{(1+\delta_{\boldsymbol{cos}})^2} & \frac{tr(\bar{\mathrm{Q}}\mathrm{K}_{\boldsymbol{sin}}\bar{\mathrm{Q}})}{(1+\delta_{\boldsymbol{sin}})^2}\end{bmatrix}\Omega\begin{bmatrix}\mathrm{y}^T\bar{\mathrm{Q}}\mathrm{K}_{cos}\bar{\mathrm{Q}}\mathrm{y} \\ \mathrm{y}^T\bar{\mathrm{Q}}\mathrm{K}_{sin}\bar{\mathrm{Q}}\mathrm{y}\end{bmatrix}$$

and

$$\Omega^{-1} \equiv \mathrm{I}_2 - \frac{N}{n}\begin{bmatrix}\frac{1}{n}\frac{tr(\bar{\mathrm{Q}}\mathrm{K}_{\boldsymbol{cos}}\bar{\mathrm{Q}}\mathrm{K}_{\boldsymbol{cos}})}{(1+\delta_{\boldsymbol{cos}})^2} & \frac{1}{n}\frac{tr(\bar{\mathrm{Q}}\mathrm{K}_{\boldsymbol{cos}}\bar{\mathrm{Q}}\mathrm{K}_{\boldsymbol{sin}})}{(1+\delta_{\boldsymbol{sin}})^2} \\ \frac{1}{n}\frac{tr(\bar{\mathrm{Q}}\mathrm{K}_{\boldsymbol{cos}}\bar{\mathrm{Q}}\mathrm{K}_{\boldsymbol{sin}})}{(1+\delta_{\boldsymbol{cos}})^2} & \frac{1}{n}\frac{tr(\bar{\mathrm{Q}}\mathrm{K}_{\boldsymbol{sin}}\bar{\mathrm{Q}}\mathrm{K}_{\boldsymbol{sin}})}{(1+\delta_{\boldsymbol{sin}})^2}\end{bmatrix}$$

Asymptotic Test Performance

Assumption (Data as concentrated random vectors)

The training data $\mathrm{x}_i \in \mathbb{R}^p, i \in \{1, \ldots, n\}$ are independently drawn from one of $K > 0$ distribution classes $\mu_1, \ldots, \mu_K$. There exist constants $C, \eta, q > 0$ such that for any $\mathrm{x}_i \sim \mu_k, k \in \{1, \ldots, K\}$ and any 1-Lipschitz function $f : \mathbb{R}^p \to \mathbb{R}$, we have

$$\mathbb{P}(|f(\mathrm{x}_i) - \mathbb{E}[f(\mathrm{x}_i)]| \geq t) \leq Ce^{-(t/\eta)^q}, \ t \geq 0$$

The test data $\hat{\mathrm{x}}_i \sim \mu_k, i \in \{1, \ldots, \hat{n}\}$ are mutually independent, but may depend on training data X and $\|\mathbb{E}[\sigma(\mathrm{WX}) - \sigma(\mathrm{W}\hat{\mathrm{X}})]\| = O(\sqrt{n})$ for $\sigma \in \{\cos, \sin\}$

Asymptotic Test Performance

Theorem (Asymptotic Test Performance)

Under Assumptions 1 and 2, we have, for E_{test} and test data $(\hat{X}, \hat{y})$ satisfying $\limsup_{\hat{n}} \|\hat{X}\|, \limsup_{\hat{n}} \|\hat{y}\| < \infty$ *with* $\hat{n}/n \in (0, \infty)$ *that, as* $n \to \infty$

$$E_{test} - \bar{E}_{test} \xrightarrow{a.s.} 0$$

$$\bar{E}_{test} = \frac{1}{\hat{n}} \|\hat{y} - \frac{N}{n} \hat{\Phi} \bar{Q} y\|^2 + \frac{N^2}{n^2 \hat{n}} \begin{bmatrix} \frac{\Theta_{cos}}{(1+\delta_{cos})^2} & \frac{\Theta_{sin}}{(1+\delta_{sin})^2} \end{bmatrix} \Omega \begin{bmatrix} y^T \bar{Q} K_{cos} \bar{Q} y \\ y^T \bar{Q} K_{sin} \bar{Q} y \end{bmatrix}$$

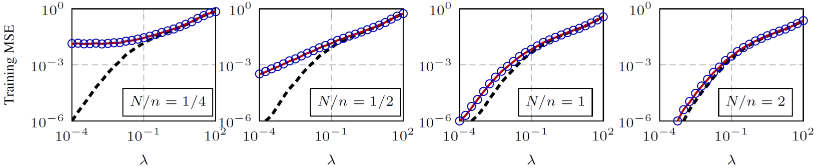
where

$$\Theta_\sigma = \frac{1}{N} tr K_\sigma(\hat{X}, \hat{X}) + \frac{N}{n} \frac{1}{n} tr \bar{Q} \hat{\Phi}^T \hat{\Phi} \bar{Q} K_\sigma - \frac{2}{n} tr \bar{Q} \hat{\Phi}^T K_\sigma(\hat{X}, X), \ \sigma \in \{cos, sin\}$$

and $\Phi \equiv \frac{K_{cos}}{1+\delta_{cos}} + \frac{K_{sin}}{1+\delta_{sin}}$, $\hat{\Phi} \equiv \frac{K_{cos}(\hat{X}, X)}{1+\delta_{cos}} + \frac{K_{sin}(\hat{X}, X)}{1+\delta_{sin}}$

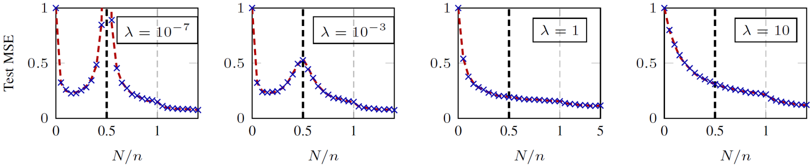
Table of Contents

Correction due to the Large n, p, M Regime



- Training MSEs of RFF ridge regression on MNIST.
  $p = 784, n = 1000, N = 250, 500, 100, 2000$ (class 3 versus 7, avg over 30 runs)
- Blue: empirical, Black: Gaussian kernel predictions ($N \to \infty$), Red: $\bar{E}_{\text{train}}$

Phase Transition and Corresponding Double Descent



- $\Omega \to \infty$ at $2N = n$ as $\lambda \to 0$
- $\bar{E}_{\text{test}} \to \infty$ as $N/n \to \frac{1}{2}$
- $\bar{E}_{\text{train}} = 0$ at $2N = n$ due to the prefactor λ^2

Table of Contents

Conclusion

- Precise description of the resolvent of RFF Gram matrices
- Asymptotic training and test performance guarantees for RFF ridge regression in the $n, p, N \to \infty$ limit
- Under- and over-parameterized regimes, involve only mild regularity assumptions on the data

References i

Belkin, Mikhail et al. (2019). "Reconciling Modern Machine Learning Practice and the Bias-Variance Trade-Off". In: arXiv: 1812.11118 [cs, stat].

Rahimi, Ali, Benjamin Recht, et al. (2007). "Random Features for Large-Scale Kernel Machines.". In: *NIPS*. Vol. 3. 4. Citeseer, p. 5.