A Brief Introduction to Double Descent

Chen Zhiyuan

Beijing Normal University

2022/4/6

Overview

1. Reconciling of Two Views

2. Simulation on RFF

3. Mathematical Analysis

4. Further Results

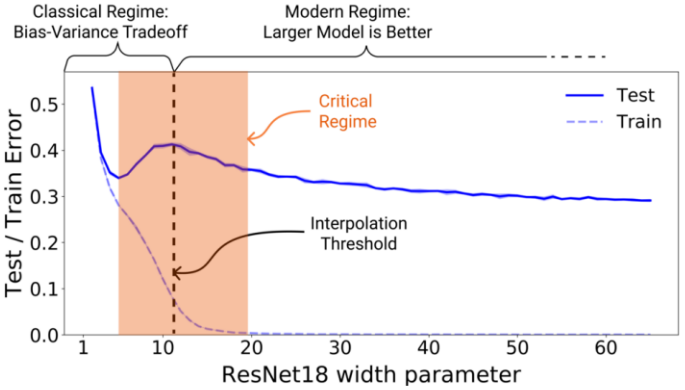
Classical and Modern Views

One of the central tenets of classical statistical learning is the bias-variance trade-off, appears to be at odds with practice in the modern machine learning practice.

The bias-variance trade-off implies that a model should balance under-fitting and over-fitting and choose a simple but effective model. Statisticians think "bigger models are worse" when beyond threshold.

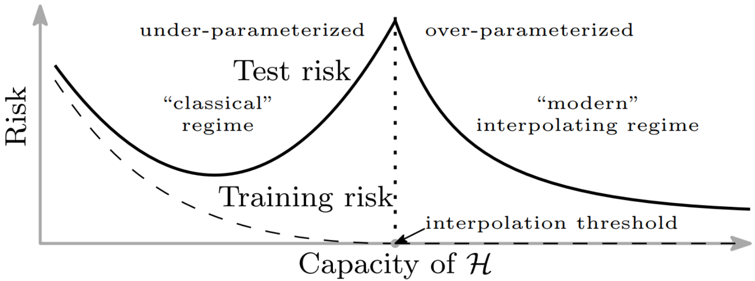
However, in modern machine learning, especially deep learning, rich models containing millions of parameters usually behave well. And relative practitioners believe that "bigger models are better".

A Suprising Phenomenon



Belkin et al.(2018) observed a phenomenon that test error decreased again after the expected U-shaped curve when the model size increased. They called this phenomenon "double descent".

Reconciling Two Views



- Classical Fields (p<n): bias-variance trade-off and U-shape curve.
- Interpolation Threshold (p=n): where training error equals to zero.
- High Dimension Fields (p>n): monotonic decline.

Random Fourier Features

Simulations are conducted through **Random Fourier Features**, which is a class of popular non-linear parametric models.

It can be viewed as a class of two-layer neural networks with fixed weights in the first layer.

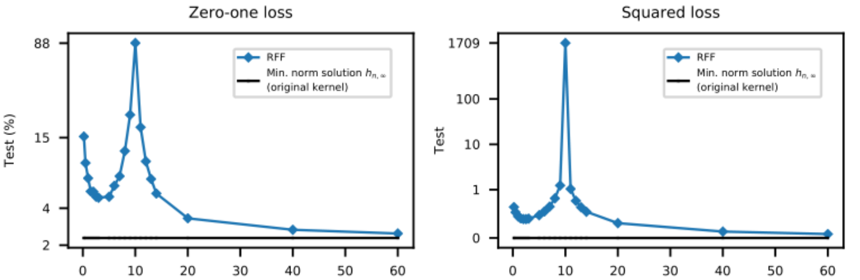
The RFF model family H_p with p parameters consists of functions $h : R^d \rightarrow C$ of the form:

$$h(x) = \sum_{k=1}^{p} a_k \phi(x; v_k) \text{ where } \phi(x; v) := e^{\sqrt{-1}<v,x>}$$

Vectors $v_1, ..., v_p$ are sampled independently from the standard normal distribution in R^d

Random Fourier Features

- Training models by minimize empirical risk objective with L_2 or L_0 loss, like $\frac{1}{n}\sum_{i=1}^{n}(h(x_i)-y_i)^2$
- When p>n, the minimizer is not unique. Choose one whose coefficients $(a_1, ..., a_p)$ have the min L_2 norm.
- Test on a subset of popular data set of handwritten digits called MNIST.



Mathematical Analysis

Belkin et al.(2020) provided a precise mathematical analysis for the shape of this curve with least square predictor.

Consider a regression problem where the response y is equal to a linear function $\beta = (\beta_1, ..., \beta_D)) \in R^D$ of D real-valued variables $X = (x_1, ..., x_D)$ plus noise $\sigma\epsilon$:

$$y = X'\beta + \sigma\epsilon$$

Given n iid copies of (x,y), we fit a linear model to the data only using a subset $T \subseteq [D] := 1, ..., D$ of $p := |T|$ variables. Then fit coefficients $\hat{\beta} = (\hat{\beta_1}, ..., \hat{\beta_D})$ with:

$$\hat{\beta}_T := (X'_T X_T)^+ X'_T y, \quad \hat{\beta}_{T^c} = 0$$

Mathematical Analysis

Theorem

Assume the distribution of X is the standard normal in R^D, ϵ is a standard normal random variable independent of X, pick any $p \in 0, ..., D$ and $T \subseteq [D]$ of cardinality p. The risk of $\hat{\beta}$ is

$$E[(y{-}X'\beta)^2] = \begin{cases} (\|\beta_{T_c}\|^2 + \sigma^2)(1 + \frac{p}{n-p-1}), & p \leq n-2 \\ +\infty, & n-1 \leq p \leq n+1 \\ \|\beta_T\|^2(1 - \frac{n}{p}) + (\|\beta_{T_c}\|^2 + \sigma^2)(1 + \frac{n}{n-p-1}), & p \geq n+2 \end{cases}$$

Random selection

proposition

If T is a uniformly random subset of [D] of cardinality p,

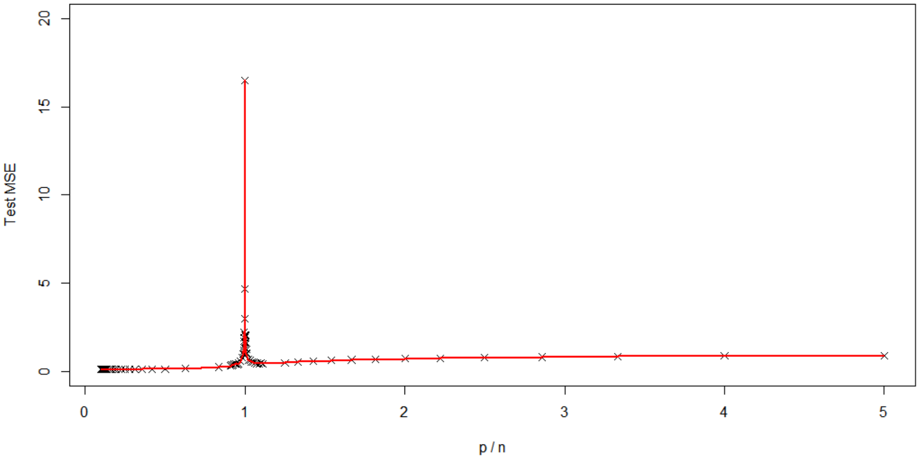
$$E[\|\beta_T\|^2] = \frac{p}{D}\|\beta\|^2, \quad E[\|\beta_{T_c}\|^2] = (1-\frac{p}{D})\|\beta\|^2$$

Corollary

Let T be a be a uniformly random subset of [D] of cardinality p. The risk of $\hat{\beta}$ is

$$E[(y - X'\beta)^2] = \begin{cases} [(1-\frac{p}{D})\|\beta\|^2 + \sigma^2](1+\frac{p}{n-p-1}), & p \leq n-2 \\ +\infty, & n-1 \leq p \leq n+1 \\ \|\beta\|^2[1-\frac{n}{D}(2-\frac{D-n-1}{p-n-1})] + \sigma^2(1+\frac{n}{n-p-1}), & p \geq n+2 \end{cases}$$

Simulation of Linear Regression



Further Results

- Nakkrian et al.(2019) conducted simulations on a wide range of neutral network models. And found not only the model size but also the training epochs will bring double descent.
- Hastie et al.(2020) found double descent occurring in fundamental models like least squares regression. And they believed when p>n, the variance decreases as p grows.

Merci