Introduction to Double Descent

Tianxiang Gao

Iowa State University

April 17, 2020
Table of Contents

1 Preliminary

2 Different types of double descent risk curve

3 Theoretical analysis for linear model in over-parameterized case
1 Preliminary

2 Different types of double descent risk curve

3 Theoretical analysis for linear model in over-parameterized case
Given a sample of training data \( S = \{(x_i, y_i)\}_{i=1}^{n} \in \mathbb{R}^p \times \mathbb{R} \).

Learn a predictor \( h : \mathbb{R}^p \to \mathbb{R} \), or denoted by \( h_n \), that is used to predict \( y \) for new \( x \).

Fix a class of functions \( \mathcal{H} \), i.e., \( h \in \mathcal{H} \), such as a neural network with a certain architecture.

Training procedure is to find a \( h \in \mathcal{H} \) that minimize the empirical risk or training error, i.e.,

\[
    h_n = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i),
\]

where \( \ell \) is a loss function, e.g., \( \ell(u, v) = \frac{1}{2} \| u - v \|^2 \).
Suppose \((x, y)\) follows some underlying distribution \(P_{xy}\), i.e., \((x, y) \sim P_{xy}\).

The \textit{generalization error} or test error for the learned predictor \(h_n\) is defined as follows

\[
R(h_n) = \mathbb{E}(\ell(h_n(x), y)) = \mathbb{E}_S \mathbb{E}_{(x, y) \sim P_{xy}} [\ell(h_n(x), y)|S]. \tag{2}
\]
Bias-Variance trade-off

The predictor class $\mathcal{H}$ is called *under-parameterized* if $p < n$.

The predictor class $\mathcal{H}$ is called *over-parameterized* if $p > n$.

In under-parameterized, minimum test risk is achieved at *sweet spot*.

The test risk is high at the interpolation threshold, where $p \approx n$.

**Figure:** Curves for training risk (dashed line) and test risk (solid line). (a) The classical U-shaped risk curve. (b) The double descent risk curve [BHMM18].
# Table of Contents

1. Preliminary

2. Different types of double descent risk curve

3. Theoretical analysis for linear model in over-parameterized case
Define the predictor class $\mathcal{H}$ as follows

$$h(x) = \sum_{k=1}^{N} a_k \phi(x; v_k), \quad \phi(x; v) = e^{\sqrt{-1} \langle x, v \rangle}$$ \hspace{1cm} (3)

where $\{v_k\}_{k=1}^{N}$ are sampled independently from a standard normal distribution on $\mathbb{R}^p$.

$\mathcal{H}$ is also denoted by $\mathcal{H}_N$, since the number $N$ is determined by a user.

Can it be considered as two-layer neural network? NO!

The features in a neural network are learned in training not fixed.

It is a linear model with nonlinear feature augmentation.

Let $z \in \mathbb{R}^N$ to be a vector with $z_k = \phi(x; v_k)$, and $Z$ denote the augmented data matrix.

The predictor is to learn $\hat{a} = \text{argmin}_a \frac{1}{2} \|Za - y\|^2$. 
Random Fourier Features [BHMM18] (continued)

**Zero-one loss**
- **Test(%)**
  - RFF
  - Min. norm solution $h_{n, \infty}$ (original kernel)

**Squared loss**
- **Test**
  - RFF
  - Min. norm solution $h_{n, \infty}$ (original kernel)

**Norm**
- **Norm**
  - RFF
  - Min. norm solution $h_{n, \infty}$

**Train(%)**
- **Train**
  - RFF

Number of Random Fourier Features ($\times 10^2$) (N)
ResNet18 [NKB+19]

Model-wise double descent.

Figure: Three layers of ResNet18 [NKB+19].
Epoch-wise double descent (in over-parameterized case).

- Small model if $p < n$.
- Intermediate model if $p \approx n$.
- Large model if $p > n$.

Figure: Test errors over model size and training epoch [NKB$^+19$].
Sample-wise double descent (in the intermediate model case).

Figure: Test error in the intermediate model case on different sample size [NKB\textsuperscript{+}19].
Table of Contents

1 Preliminary

2 Different types of double descent risk curve

3 Theoretical analysis for linear model in over-parameterized case
Data model and risk [HMRT19]

- Assume the observed sample \( \{(x_i, y_i)\}_{i=1}^n \in \mathbb{R}^p \times \mathbb{R} \) from a model

\[
(x_i, \epsilon_i) \sim P_x \times P_\epsilon
\]

\[
y_i = x_i^T \beta + \epsilon_i
\]  

where the random draws across \( i = 1, \ldots, n \) are independent. Here, \( P_x \) is a distribution on \( \mathbb{R}^p \) such that \( \mathbb{E}(x_i) = \mu \) and \( \text{Cov}(x_i) = \Sigma \), and \( P_\epsilon \) is a distribution on \( \mathbb{R} \) with \( \mathbb{E}(\epsilon_i) = 0 \) and \( \text{Var}(\epsilon_i) = \sigma^2 \).

- The least-norm solution \( \hat{\beta} = \arg\min \{ \| b \| : b \text{ minimize } \| y - Xb \|^2 \} \)

- \( \hat{\beta} = \lim_{\lambda \to 0^+} \hat{\beta}_\lambda \), where

\[
\hat{\beta}_\lambda = \arg\min_b \frac{1}{n} \| y - Xb \|^2 + \lambda \| b \|^2
\]
The least-norm solution is obtained by gradient descent.

**Classical Results**

The least-norm solution is obtained by gradient descent.

**[HMRT19, Proposition 1]**

Initialize $\beta^0 = 0$, and consider running gradient descent on the least square loss, yielding iterates

$$
\beta^{k+1} = \beta^k - tX^T(X\beta^k - y),
$$

(7)

where $0 < t \leq 1/\lambda_{\max}X^T X$. Then $\lim_{k \to \infty} \beta^k = \hat{\beta}$.
Classical Results

The Bias and Variance of the min-norm solution.

\[ R_X(\hat{\beta}; \beta) = \frac{1}{\|X\|^2} \left( \|\mathbb{E}(\hat{\beta} | X) - \beta\|_\Sigma^2 + \text{tr} [\text{Cov}(\hat{\beta} | X) \Sigma] \right). \]

\[ B_X(\hat{\beta}; \beta) \quad \text{and} \quad V_X(\hat{\beta}; \beta) \]

**Lemma 1.** Under the model (2), (3), the min-norm least squares estimator (4) has bias and variance

\[ B_X(\hat{\beta}; \beta) = \beta^T \Pi \Sigma \Pi \beta \quad \text{and} \quad V_X(\hat{\beta}; \beta) = \frac{\sigma^2}{n} \text{tr}(\hat{\Sigma}^+ \Sigma), \]

where \( \hat{\Sigma} = X^T X / n \) is the (uncentered) sample covariance of \( X \), and \( \Pi = I - \hat{\Sigma}^+ \hat{\Sigma} \) is the projection onto the null space of \( X \).

**Theorem 1.** Assume the model (2), (3), and assume \( x \sim P_x \) is of the form \( x = \Sigma^{1/2} z \), where \( z \) is a random vector with i.i.d. entries that have zero mean, unit variance, and a finite 4th moment, and \( \Sigma \) is a deterministic positive definite matrix, such that \( \lambda_{\min}(\Sigma) \geq c > 0 \), for all \( n, p \) and a constant \( c \) (here \( \lambda_{\min}(\Sigma) \) is the smallest eigenvalue of \( \Sigma \)). As \( n, p \to \infty \), assume that the spectral distribution \( F_\Sigma \) converges weakly to a measure \( H \). Then as \( n, p \to \infty \), such that \( p/n \to \gamma < 1 \), the risk of the least squares estimator (4) satisfies, almost surely,

\[ R_X(\hat{\beta}; \beta) \to \sigma^2 \frac{\gamma}{1 - \gamma}. \]
Theorem 2. Assume the model (2), (3), where $x \sim P_x$ has i.i.d. entries with zero mean, unit variance, and a finite moment of order $8 + \eta$, for some $\eta > 0$. Also assume that $\|\beta\|_2^2 = r^2$ for all $n, p$. Then for the min-norm least squares estimator $\hat{\beta}$ in (4), as $n, p \to \infty$, such that $p/n \to \gamma > 1$, it holds almost surely that

$$R_X(\hat{\beta}; \beta) \to r^2(1 - 1/\gamma) + \frac{\sigma^2}{\gamma - 1}.$$
Questions?
References

