Double Descent in a Linear Model

Ngo Cuong
Part C Student
Department of Statistics
University of Oxford

11 March 2020

Oxford

OVERVIEW

- "Suprises in High-Dimensional Ridgeless Least Squares Interpolation" by Trevor Hastie et al., 2019.
- Reconciling the classical Bias-Variance trade-off. Consequences of interpolation.
- Double Descent curve in a linear setting.

Supervised Learning

Input: (x_i, y_i) for $i = 1, ..., n$. Aim find $f : \mathbb{R}^p \to \mathbb{R}$ such that $f(x)$ match y for test data (x, y).

Goal of Machine Learning:
$f_{opt} = argmin_f \mathbb{E}_{unseen\,data} L(f(x), y)$

In practice we solve ERM:
$f_{ERM} = argmin_{f \in \mathbb{H}} \frac{1}{n} \sum_{training\,data} L(f(x_i), y_i)$

Note:
$\mathbb{E}_{unseen\,data} L(f(x), y) \leq \frac{1}{n} \sum_{training\,data} L(f(x_i), y_i) + O(\sqrt{\frac{c}{n}})$

Risk curve

The traditional understanding of the Bias-Variance trade-off:

- Classical U-shaped curve.

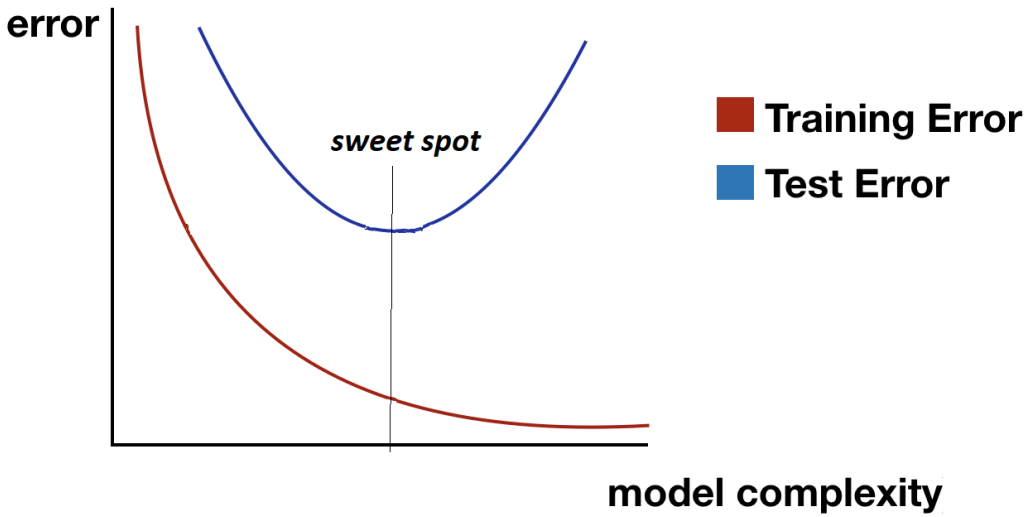


Figure 1: The classical U curve of the generalisation/test error.

Risk curve

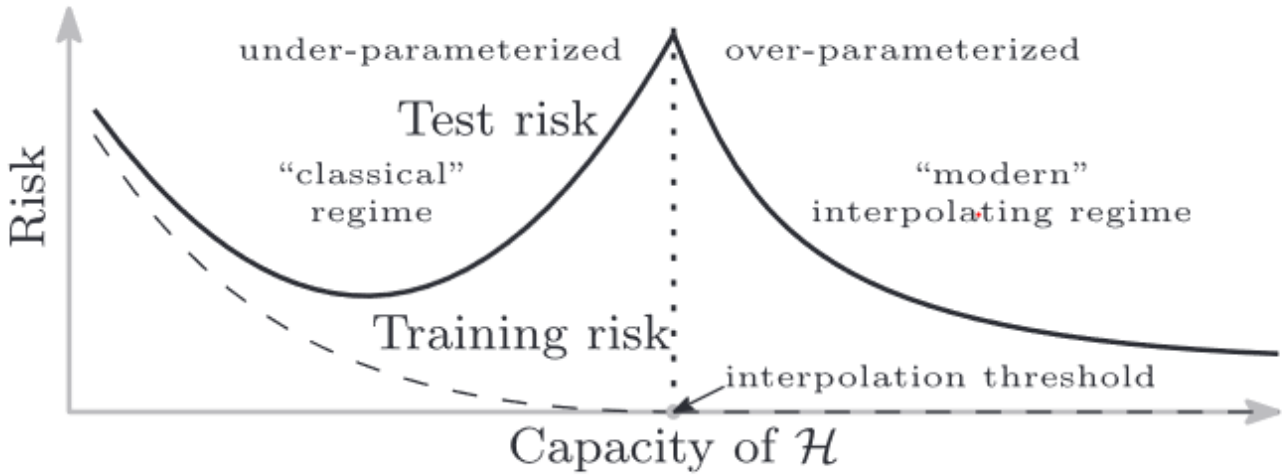
model	# params	random crop	weight decay	train accuracy	test accuracy
Inception	1,649,402	yes	yes	100.0	89.05
		yes	no	100.0	89.31
		no	yes	100.0	86.03
		no	no	100.0	85.75

CIFAR 10; Understanding deep learning requires rethinking generalization, Zhang et al., 2017.

Risk curve

The modern approach to the Bias-Variance trade-off :

- Double descent curve.
- "Interpolation does not contradict generalization".



Reconciling modern machine learning practice and the bias-variance trade-off, Belkin et al., 2018

Linear model setup

Suppose we have n independent and identically distributed training samples $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}, i = 1, 2...n$, where each sample is generated independently from the following procedure (1):

- Draw $x_i \sim F_x$ and $\epsilon_i \sim F_\epsilon$ independently from some probability distributions F_x and F_ϵ.
- Obtain $y_i = x_i^T \beta + \epsilon_i$ for some $\beta \in \mathbb{R}^p$.
- Assume ϵ_i is independent homoscedastic noise such that $\mathbb{E}(\epsilon_i) = 0$ and $\mathbb{E}(\epsilon_i^2) = \sigma^2$.
- Assume feature vectors x_i have mean $\mathbb{E}(x_i) = 0$, have covariance matrix $cov(x_i) = \Sigma$.

Risk function

Define the *sample prediction risk* for a *fresh unseen* data sample $x \sim F_x$:

$$
\begin{aligned}
R_{\boldsymbol{X}}(\hat{\beta}) &= \mathbb{E}[(x^T\hat{\beta} - x^T\beta)^2|\boldsymbol{X}] \\
&= \mathbb{E}[(x^T(\hat{\beta} - \beta)^2)|\boldsymbol{X}] \\
&= \mathbb{E}[Tr((\hat{\beta} - \beta)x^T x(\hat{\beta} - \beta)^T)|\boldsymbol{X}] \\
&= \mathbb{E}[(\hat{\beta} - \beta)\Sigma(\hat{\beta} - \beta)^T|\boldsymbol{X}] \\
&= \mathbb{E}[||\hat{\beta} - \beta||_\Sigma^2|X]
\end{aligned}
$$

We consider the least squares estimator with respect to the $L2$-norm: $\hat{\beta} = (X^TX)^\dagger X^T y$ where $(X^TX)^\dagger$ is Moore Penrose pseudo-inverse of X^TX.

Risk function

We can derive the Bias-Variance decomposition of the risk:

$$R_X(\hat{\beta};\beta) = \underbrace{\|\mathbb{E}(\hat{\beta}|X) - \beta\|_\Sigma^2}_{B_X(\hat{\beta};\beta)} + \underbrace{\mathrm{tr}[\mathrm{Cov}(\hat{\beta}|X)\Sigma]}_{V_X(\hat{\beta};\beta)}.$$

Further we obtain:

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

where $\hat{\Sigma} = \frac{X^T X}{n}$ is the sample covariance of X, and $\Pi = \mathsf{I} - \hat{\Sigma}^\dagger\hat{\Sigma}$ is the projection onto the nullspace of X.

Asymptotic Risk

We set the dimension of the features and the number of observations to go to infinity in a proportional regime $p/n \to \gamma$ as $n, p \to \infty$, where $\gamma \in (0, \infty)$. We measure "model complexity" with respect to γ.
Underparametrized regime is when $\boldsymbol{\gamma < 1}$.
Overparametrized regime is when $\boldsymbol{\gamma > 1}$.
Interpolation threshold is when $\boldsymbol{\gamma = 1}$.

Main Theorem

Theorem. Consider the linear model setup described above with signal $||\beta||_2^2 = r^2$. Then when $p/n \to \gamma$ as $p, n \to \infty$, we have almost surely:

$$R_X(\widehat{\beta}) \to \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma} & \text{if } \gamma < 1 \\ \sigma^2 \frac{1}{\gamma - 1} + r^2\left(1 - \frac{1}{\gamma}\right) & \text{if } \gamma > 1 \end{cases}$$

In the underparametrized regime the risk is pure variance, whereas in the overparametrized regime the risk is a sum of bias and variance terms.

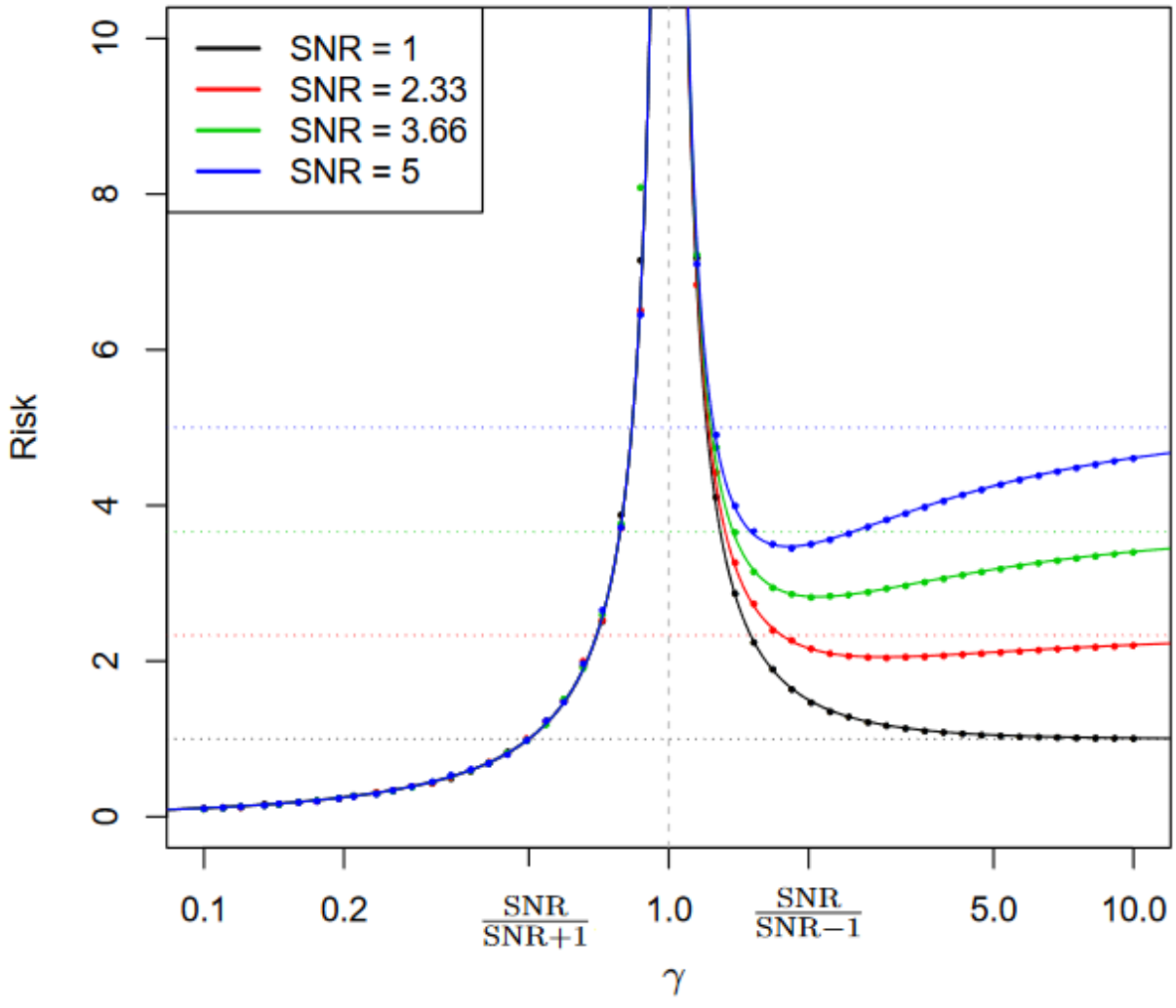
Theorem Insights

$$R_X(\widehat{\boldsymbol{\beta}}) \to \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma} & \text{if } \gamma < 1 \\ \sigma^2 \frac{1}{\gamma - 1} + r^2\left(1 - \frac{1}{\gamma}\right) & \text{if } \gamma > 1 \end{cases}$$

Define the null risk to be the loss when $\hat{\beta} = 0$.

- The null risk is equal to r^2. $\mathbb{E}[||\hat{\beta} - \beta||_{\Sigma}^2 | X] = r^2$
- The two cases $\gamma < 1$ and $\gamma > 1$ align as γ approaches 1.
- In the underparametrized regime the least squares risk $R_X(\gamma)$ is better than the null risk iff $\gamma < \frac{SNR}{SNR+1}$.

Theorem Insights



Theorem Insights

- In the overparametrized regime when SNR ≤ 1, the least squares risk is always worse than the null risk. Moreover it is monotonically deacreasing and approaches the null risk from above as $\gamma \to \infty$.
- In the overparametrized regime when SNR > 1, the least squares risk is better than the null risk iff $\gamma > \frac{SNR}{SNR-1}$. Not monotonically deacreasing but with local minimum at $\gamma = \frac{\sqrt{SNR}}{\sqrt{SNR}-1}$ and approaches the null risk from below as $\gamma \to \infty$.

Theorem Assumptions for underparametrized regime

Key assumptions:

- The feature vector x is of the form $x = \Sigma^{1/2} z$, where z is a random vector with i.i.d entries with zero mean and unit variance.
- Σ is a deterministic positive definite matrix, i.e. has strictly positive eigenvalues.
- $p/n \to \gamma < 1$, as $n, p \to \infty$.

Underparametrized regime

Recall the risk function has the following form:

$$R_X(\hat{\beta};\beta) = \underbrace{\|\mathbb{E}(\hat{\beta}|X) - \beta\|_\Sigma^2}_{B_X(\hat{\beta};\beta)} + \underbrace{\mathrm{tr}[\mathrm{Cov}(\hat{\beta}|X)\Sigma]}_{V_X(\hat{\beta};\beta)}.$$

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

where $\hat{\Sigma} = \frac{X^T X}{n}$ is the sample covariance of X, and $\Pi = \mathsf{I} - \hat{\Sigma}^\dagger\hat{\Sigma}$ is the projection onto the nullspace of X.

Proof Outline

Step 1: Bias is almost surely zero. Show sample covariance $\hat{\Sigma}$ is almost surely invertible.

$$\lambda_{\min}(X^T X/n) \geq \lambda_{\min}(Z^T Z/n)\lambda_{\min}(\Sigma) \geq (c/2)(1-\sqrt{\gamma})^2,$$

Step 2: Write the variance with respect to the spectral measure of $Z^T Z/n$.

$$V_X(\hat{\beta};\beta) = \frac{\sigma^2 p}{n}\int \frac{1}{s}\, dF_{Z^T Z/n}(s)$$

Proof Outline

Step 3: Apply Marcheko-Pastur convergence theorem.

$$V_X(\hat{\beta};\beta) \to \sigma^2\gamma \int \frac{1}{s}\, dF_\gamma(s).$$

Step 4: Stieltjes transform $m(z)$ of Marchenko-Pastur law at $z = 0$.

$$\int \frac{1}{s}\, dF_\gamma(s) = m(-z)\Big|_{z=0} = \frac{-(1-\gamma+z)+\sqrt{(1-\gamma+z)^2+4\gamma z}}{2\gamma z}\Bigg|_{z=0} = \frac{1}{1-\gamma}$$

What's next?

- Overparameterized regime - Variance is derived using the same approach as underparameterized regime. Bias is not zero anymore.
- Gaussian features - quick way to derive the bias.
- Isotropic features - needs generalized Marchenko-Pastur theorem.
- Non-Linear setting.