Lecture Notes VII – Double descent

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

December, 2021

“Double descent” on a simple example

Reading HTF Ch.: , Murphy Ch.: , Bach Ch.: Ch.10.2.3

Linear regression when $n > d$

- We describe a very simple linear regression situation (following Bach section 10.2.3)
- For it, we are able to explicitly obtain the expected estimation error $E[\|\theta_{true} - \hat{\theta}\|^2]$
- Surprisingly, the variance of this error decreases with d, and the error itself has a limit proportional to $\|\theta_{true}\|^2$.

- Input distribution $x^{1:n} \sim N(0, I_d)$, noise $\epsilon^{1:n} \sim N(0, \sigma^2)$
- Model $y^i = (x^i)^T \theta_{true} + \epsilon^i$.
- Denote $X \in \mathbb{R}^{n \times d}$, $y, \epsilon \in \mathbb{R}^n$ the usual input matrix, output, and noise vectors respectively
- Denote $K = XX^T \in \mathbb{R}^{n \times n}$ the Gram matrix (or kernel matrix). We assume K is non-singular

- From Lecture IV, The Implicit Bias of Gradient Descent we know that
  - When X is full rank n, the equation $y = X\theta$ has multiple solutions θ
  - Gradient Descent converges to the min norm solution $\hat{\theta} = X^T K^{-1} y$

The expected estimation error $MSE(\theta_{true}) = E[\|\theta_{true} - \hat{\theta}\|^2]$

- Decompose $\hat{\theta}$

$$\hat{\theta} \;=\; X^T K^{-1} y \;=\; X^T K^{-1}(X\theta_{true} + \epsilon) \;=\; X^T K^{-1} X \theta_{true} + X^T K^{-1} \epsilon \tag{1}$$

- Then,

$$\begin{aligned} MSE(\theta_{true}) &= E_{X,\epsilon}[\|\theta_{true} - \hat{\theta}\|^2] && (2) \\ &= \underbrace{E_X[\theta_{true}^T (I_d - X^T K^{-1} X)\theta_{true}]}_{\text{bias}^2} + \underbrace{E_{X,\epsilon}[\epsilon^T K^{-1} X X^T K^{-1} \epsilon]}_{\text{variance}} && (3) \end{aligned}$$

- The Variance term becomes

$$\begin{aligned} Var &= E_{X,\epsilon}[\epsilon^T K^{-1} \epsilon] && (4) \\ &= E_X[\operatorname{trace} K^{-1}]\sigma^2 \qquad \text{Wishart!} && (5) \\ &= \sigma^2 \frac{n}{d - n - 1} && (6) \end{aligned}$$

The expected estimation error $MSE(\theta_{true}) = E[\|\theta_{true} - \hat{\theta}\|^2]$

- The Bias^2 term:
- Note that $\theta_P = X^T K^{-1} X \theta_{true}$ is the orthogonal projection of θ_{true} on the row space of X, and $\theta_{true}^T X^T K^{-1} X)\theta_{true} = \|\theta_P\|^2$.
- The subspace is a random subspace of dimention n in $\mathbb{R}^d$. By spherical symmatry, the length of the projection of a fixed vector on a random subspace is the same with that of a projecting a random vector of length (squared) $\|\theta_{true}\|^2$ on a fixed subspace, e.g. the first d unit vectors in $\mathbb{R}^d$. The latter expected value is easy to compute

$$E[\|\theta_P\|^2] \; = \; \frac{n}{d}\|\theta_{true}\|^2 \tag{7}$$

  Exercise Proving this is a moderatly easy exercise
- Hence,

$$\text{bias}^2 \; = \; E_X[\theta_{true}^T(I_d - X^T K^{-1} X)\theta_{true}] \; = \; \frac{d-n}{d}\|\theta_{true}\|^2 \tag{8}$$

The expected estimation error $MSE(\theta_{true}) = E[\|\theta_{true} - \hat{\theta}\|^2]$

- Finally
$$MSE(\theta_{true}) \;=\; E[\|\theta_{true} - \hat{\theta}\|^2] \;=\; \frac{d-n}{d}\|\theta_{true}\|^2 + \sigma^2 \frac{n}{d-n-1} \tag{9}$$
for $d > n+1$
- When $d \to \infty$, the variance$\to 0$ and the bias$^2 \to \|\theta_{true}\|^2$