Bias Variance Decomposition and Double descent : first part.

The definition of PAC learning indicates that to achieve good generalization error $L_{\mathcal{D}}(A(S))$ or its expected value $\mathbb{E}_S[L_{\mathcal{D}}(A(S))]$ at least two aspects come into play : the choice of the learning rule $A(S)$ and the choice of the hypothesis class $\mathcal{H} \ni A(S)$. For the choice of $\mathcal{H}$, in particular, a large enough $\mathcal{H}$ is desirable as it makes $\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ small and $L_{\mathcal{D}}(A(S))$ potentially small, but on the other hand a too large $\mathcal{H}$ is detrimental because of the No free Lunch Theorem.

The right choice of $\mathcal{H}$ turns out be a subtle

Issue. According to the classical paradigm of statistics it should be sufficiently complex as to be expressive enough but not too large at the same time. This is often expressed as the so-called bias-variance tradeoff. However the classical bias-variance tradeoff has been challenged in recent years as exemplified by the "double descent phenomenon".

In this lecture and the next one, after reviewing the bias-variance trade-off we look at the double descent phenomenon. We will look at it through a nice simple variant of a regression model first introduced by Breiman and Freedman 1983 and revisited by Belkin, Hsu, Xu in 2018.

1. Bias-Variance Tradeoff : classical approach.

- Let S = training set = $\{(x_1, y_1) \dots (x_m, y_m)\}$
  $x \in \mathbb{R}^d$, $y \in \mathbb{R}$.
- Square loss $\ell(x, y) = (h(x) - y)^2$
  $= \ell(h; x, y)$.
- Population risk $L_{\mathcal{D}}(h) = \mathbb{E}_{\mathcal{D}}(\ell(h; x, y))$
  $= \mathbb{E}_{\mathcal{D}}((h(x) - y)^2)$
- Learning rule or predictor : $A(S) = h_S$.
- Expected test loss of learning rule :

$$\mathbb{E}_S(L_{\mathcal{D}}(h_S)).$$

is the average over training sets of error of learning rule $L_{\mathcal{D}}(h_S)$. This will be our main quantity of interest today.

Lemma.

$$\min L_{\mathcal{D}}(h) = \min \mathbb{E}\left((h(x) - y)^2\right)$$

is attained by the "optimal learning rule"

$$h_{opt}(x) = \mathbb{E}(y|x) = \int_{\mathbb{R}} dy\, y\, \mathcal{D}(y|x)$$

Proof

$$\mathbb{E}_{\mathcal{D}}\left((h(x)-y)^2\right) = \mathbb{E}\left((h(x)-h_{opt}(x) + h_{opt}(x)-y)^2\right)$$

$$= \mathbb{E}\left(h(x)-h_{opt}(x)\right)^2 + \mathbb{E}\left(h_{opt}(x)-y\right)^2 + 2\,\mathbb{E}\left[(h(x)-h_{opt}(x))(h_{opt}(x)-y)\right]$$

Now $\mathbb{E}_{\mathcal{D}} = \mathbb{E}_x \mathbb{E}_{y|x}$ (since $\mathcal{D}(x,y) = \mathcal{D}(y|x)\mathcal{D}(x)$)

Thus the third expectation is:

$$\mathbb{E}_x\Big[(h(x) - h_{opt}(x))\underbrace{\mathbb{E}_{y|x}(h_{opt}(x) - y)}_{h_{opt}(x) - \mathbb{E}_{y|x}(y)}\Big]$$

$$\underbrace{\qquad\qquad}_{0}.$$

Therefore

$$L_{\mathcal{D}}(h) = \mathbb{E}_{\mathcal{D}}(h(x) - y)^2$$

$$= \underbrace{\mathbb{E}(h(x) - h_{opt}(x))^2}_{\text{error induced by choice of rule } h \text{ as compared to optimal rule.}} + \underbrace{\mathbb{E}\big((\mathbb{E}(y|x) - y)^2\big)}_{\text{structural noise term here to remain}}$$

In particular:

$$L_{\mathcal{D}}(h) \geq \text{Noise term} = \mathbb{E}\big((\mathbb{E}(y|x) - y)^2\big)$$

with equality attained for $h = h_{opt}$

∎

Example of Noise term.

Imagine data S as follows for $i = 1 \dots m$

$$y_i = \vec{\beta}_*^T \cdot \vec{x}_i + \varepsilon_i$$

$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $\vec{\beta}_* \in \mathbb{R}^d$ (unknown or hidden ground truth).

Here $\mathcal{D}(x, y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \vec{\beta}_* \cdot \vec{x})^2}{2\sigma^2}\right)$

Thus

$$h_{opt}(x) = \mathbb{E}(y \mid x) = \int dy \; y \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y - \vec{\beta}_* \cdot \vec{x})^2}{\sigma^2}}$$

$$\Rightarrow \quad h_{opt}(x) = \vec{\beta}_* \cdot \vec{x}$$

$$L_{\mathcal{D}}(h) = \mathbb{E}\left((h(x) - \vec{\beta}_* \cdot \vec{x})^2\right)$$

$$+ \underbrace{\mathbb{E}\left((y - \vec{\beta}_* \vec{x})^2\right)}_{\mathbb{E}(\varepsilon^2) = \sigma^2} \leftarrow \text{Noise}$$

Imagine furthermore $h(x) = \hat{\beta} \cdot \vec{x}$ for a linear model.

$$\mathbb{E}(h(x) - \hat{\beta}_* \vec{x})^2 = \mathbb{E}((\hat{\beta} - \beta_*)^T \cdot \vec{x})^2$$

$$= \mathbb{E}\left[(\hat{\beta} - \beta_*)^T \underbrace{\vec{x}\,\vec{x}^T}(\hat{\beta} - \beta_*)\right]$$

$$= (\hat{\beta} - \beta_*^T)\,\Sigma\,(\hat{\beta} - \beta_*)$$

with $\Sigma = \mathbb{E}(\vec{x}\,\vec{x}^T)$ = Covariance matrix of data. (unknown if data distr is unknown.).

Theorem : bias - variance decomposition.

$$\mathbb{E}_S\left[L_{\mathcal{D}}(h_S)\right]$$

$$= \mathbb{E}_{\mathcal{D}_X}\left[\left(\mathbb{E}_S(h_S(x)) - h_{opt}(x)\right)^2\right] \quad \text{bias of } h_S(x).$$

$$+ \mathbb{E}_S\,\mathbb{E}_{\mathcal{D}_X}\left[\left(h_S(x) - \mathbb{E}_S(h_S(x))\right)^2\right] \quad \text{Variance of } h_S(x)$$

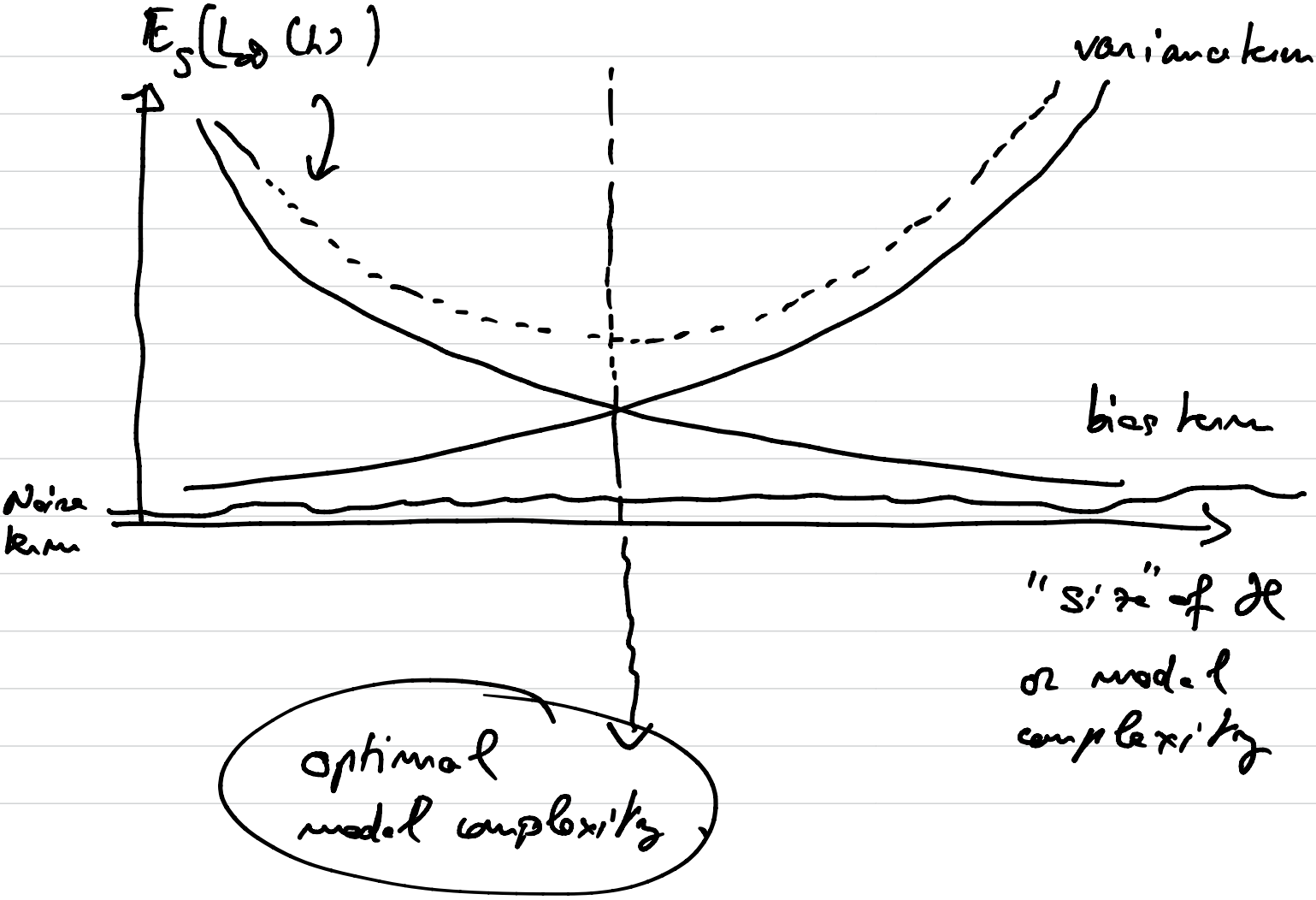
$$+ \mathbb{E}_{\mathcal{D}}\left[\left(h_{opt}(\vec{x}) - y\right)^2\right] \quad \text{Noise}$$

Remarks:

- Third term is the incompressible structural noise term here even for the optimal rule.
- First term is called bias term. It is the error incurred by the learning rule w.r.t the optimum estimator. For very expressive classes of this tends to be small because

$h_S(x)$ might be close to $h_{opt}(x)$.

- Second term is called variance term. This term quantifies the variability of estimator when the data set changes. Naively speaking if the class $\mathcal{H}$ is too expressive we might overfit to the particular data set (e.g we might even fit noise) so that the variance grows.



Proof.

Recall $\mathbb{E}_S(L_{\mathcal{D}}(h_S)) = \mathbb{E}_S\,\mathbb{E}_{\mathcal{D}}\left[(h_S(x)-y)^2\right]$

$$\mathbb{E}_{\mathcal{D}}(h_S(x)-y)^2 = \mathbb{E}_{\mathcal{D}}(h_S(x)-h_{opt}(x)+h_{opt}(x)-y)^2$$

$$= \mathbb{E}_{\mathcal{D}}(h_S(x)-h_{opt}(x))^2 + \mathbb{E}(h_{opt}(x)-y)^2$$

$$+ 2\,\mathbb{E}_{\mathcal{D}}(h_S(x)-h_{opt}(x))(h_{opt}(x)-y)$$

$$\uparrow$$

$$\mathbb{E}_{\mathcal{D}_x}\,\mathbb{E}_{\mathcal{D}_{y|x}}$$

using the $\mathbb{E}_{\mathcal{D}_{y|x}}(h_{opt}(x)-y)$

$$= h_{opt}(x) - \mathbb{E}_{\mathcal{D}_{y|x}}(y) = 0$$

the cross term disappears. Thus

$$\mathbb{E}_S(L_{\mathcal{D}}(h_S)) = \mathbb{E}_S\,\mathbb{E}_{\mathcal{D}}(h_S(x)-h_{opt}(x))^2$$

$$+ \mathbb{E}(h_{opt}(x)-y)^2 \leftarrow \text{Noise term}$$

It remains to decompose the first term:

$$\mathbb{E}_S \mathbb{E}_{\mathcal{D}} \left(h_S(x) - h_{opt}(x)\right)^2$$

$$= \mathbb{E}_S \mathbb{E}_{\mathcal{D}} \left(h_S(x) - \mathbb{E}(h_S(x))\right)^2$$

$$+ \mathbb{E}_S \mathbb{E}_{\mathcal{D}} \left(\mathbb{E}(h_S(x)) - h_{opt}(x)\right)^2$$

$$+ 2\mathbb{E}_S \mathbb{E}_{\mathcal{D}} \left(h_S(x) - \mathbb{E}(h_S(x))\right)\left(\mathbb{E}(h_S(x)) - h_{opt}(x)\right)$$

Ucinz $\mathbb{E}_S \left(h_S(x) - \mathbb{E}(h_S(x))\right) = 0$ & that the second term is indep of S:

$$\mathbb{E}_S \mathbb{E}_{\mathcal{D}} \left(h_S(x) - h_{opt}(x)\right)^2$$

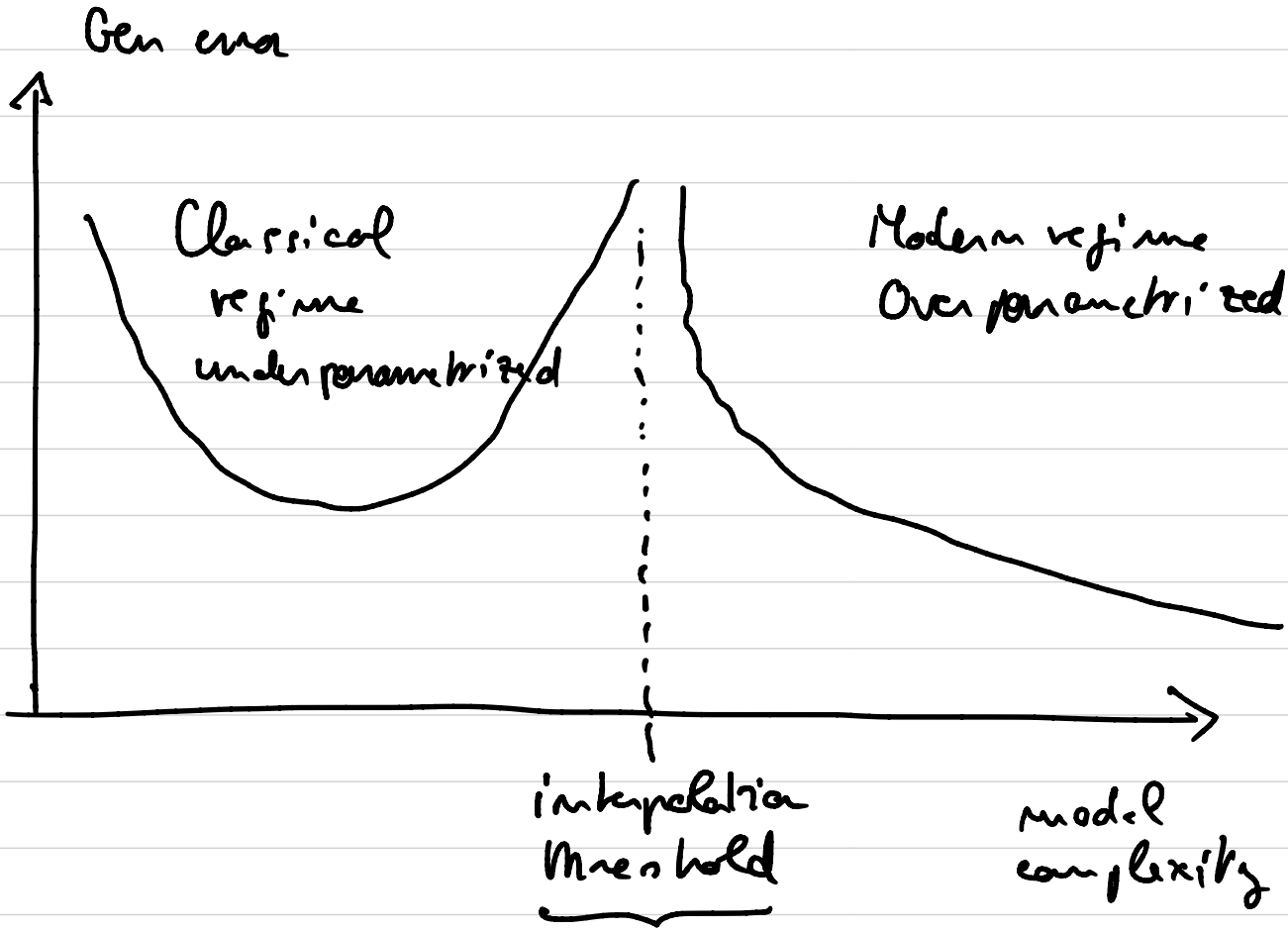
$$= \mathbb{E}_{S,\mathcal{D}_x} \left(h_S(x) - \mathbb{E}(h_S(x))\right)^2 \leftarrow \text{variance}$$

$$+ \mathbb{E}_{\mathcal{D}_x} \left(\mathbb{E}(h_S(x)) - h_{opt}(x)\right)^2 \leftarrow \text{bias}$$

■

2. Double descent phenomenon.

The heuristic picture of the generalization curve on page 9 has been challenged. Next week we will see on a specific model (a variant of regression) what one can have:



model complexity = number of data samples.

Understanding this phenomenon and when it happens is an important problem. This kind of curve can also be richer with many peaks occurring.

Model of Breiman-Freedman / revisited by Belkin, Xsu, Hu.

We assume that in the data $(\vec{x}_1, y_1) \dots (\vec{x}_m, y_m)$ only the components $j \in A$, $A \subset \{1 \dots d\}$ of $\vec{x}_j$ are observed. We set $\vec{x}_1^A, \vec{x}_2^A \dots \vec{x}_m^A$ for the vectors $\vec{v}^A = [v_j, j \in A]$ with $|A| = p \leq d$ which are d-dimensional.

Moreover $y_i = \vec{\beta}_*^T \cdot x_i + \varepsilon_i$, $\varepsilon_i \in \mathcal{N}(0,1)$.

Prediction rule from linear class $\mathcal{H}$: $h_S(x) = \vec{\beta}_A^T \cdot \vec{x}_A$

$\vec{\beta}_A \in \mathbb{R}^d$ where it minimizes the empirical loss

$$\frac{1}{m} \sum_{i=1}^{m} (y_i - \vec{\beta}_A^T \cdot \vec{x}_A^i)^2 = L_S(\beta_A).$$

Notation:

$$X_A = \begin{bmatrix} x_A^{1\,T} \\ \vdots \\ x_A^{m\,T} \end{bmatrix} \in \mathbb{R}^{m \times p} \qquad p \leq d$$

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} \in \mathbb{R}^m$$

$$\left[\vec{\beta}_A^{\,T} \cdot \vec{x}_A^{\,i} \right] = \underbrace{\left[\vec{x}_A^{\,i\,T} \cdot \vec{\beta}_A \right]}_{\substack{m\text{-dim vector} \\ i = 1 \dots m}} = \underbrace{X_A \vec{\beta}_A}_{\substack{\mathbb{R}^{m \times p} \times \mathbb{R}^p \\ \mathbb{R}^m}}$$

$$L_S(\vec{\beta}_A) = \frac{1}{m} \| Y - X_A \beta_A^T \|^2$$

$$\text{Let } \hat{\beta} = \begin{cases} \hat{\beta}_A & \text{minimizer of } L_S(\vec{\beta}_A) \text{ for } j \in A \\ 0 & \text{for } j \in A^c. \end{cases}$$

$$L_{\mathcal{D}}(\vec{\beta}) = \mathbb{E}_{\mathcal{D}} \left[(y - \hat{\beta} \cdot \vec{x})^2 \right]$$

Recap next time

3. Mathematical recap on least squares minimization.

Let $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix} \in \mathbb{R}^{m \times d}$ matrix.

Let $Y \in \mathbb{R}^m$ vector

Let $L(\vec{\beta}) = \|Y - X\vec{\beta}\|^2$

$$= \sum_{i=1}^{m} (y_i - x_i^T \cdot \vec{\beta})^2$$

$$= \sum_{i=1}^{m} (y_i - \vec{\beta}^T \cdot \vec{x}_i)^2 .$$

We want to minimize $L(\vec{\beta})$. The naive approach is to compute $\frac{\partial}{\partial \beta_k} L(\vec{\beta}) = 0$ i.e:

$$\sum_{i=1}^{m} 2(y_i - \vec{\beta}^T \cdot \vec{x}_i)\, x_{ik} = 0 \quad , \quad k = 1 \cdots d$$

$$\Rightarrow \sum_{i=1}^{m} x_{ik} y_i = \sum_{\ell=1}^{d} \beta_\ell \sum_{i=1}^{m} x_{i\ell} x_{ik}$$

$$\Rightarrow \boxed{X^T Y = (X^T X)\vec{\beta}} \quad (*)$$

How do we solve this equation ?

a) $\underline{\underline{d \leq n}}$ $\quad X^T X \in \mathbb{R}^{d \times d}$ and generically this matrix is full rank (we assume it here).

so it is invertible and the solution is unique.

$$\boxed{\vec{\beta}_{LS} = (X^T X)^{-1} X^T Y.}$$

Nek $L(\vec{\beta}_{LS}) = \| Y - X(X^T X)^{-1} X^T Y \|^2$

$= \| (\mathbb{1} - X(X^T X)^{-1} X^T) Y \|^2$

- $X (X^T X)^{-1} X^T = P$ is symmetric $P = P^T$
- $P^2 = X(X^T X)^{-1} X^T X (X^T X)^{-1} X^T = P$

{ so this is a projection matrix: it take vector $Y \in \mathbb{R}^n$ and projects it on d-dim subspace spanned by rows of X^T.
And $(\mathbb{1} - P)$ is orth proj

b) $\underline{d > n}$ $\quad X X^T \in \mathbb{R}^{n \times n}$ is full rank and invertible generically. But $X^T X$ has $d - n$ zero eigenvalues.

Recall equation (*):

$$X^T Y = (X^T X)\beta$$

set $\beta = \underbrace{\beta_{particular}}_{X^T (X X^T)^{-1} Y} + \underbrace{z}_{\text{vector in null space of } X^T X \text{ (a linear combination of zero eigvects)}}$

we have

$$(X^T X)\beta = \underbrace{X^T X \, X^T (X X^T)^{-1} Y}_{X^T Y} + \underbrace{(X^T X) z}_{0}$$

so the equation is satisfied.

We see there is a whole set of solutions:

$$\boxed{\beta_{LS} = X^T (X X^T)^{-1} Y + z, \quad z \in Null(X^T X)}$$

Note $L(\beta_{LS}) = \|Y - X\beta\|^2 = \|Xz\|^2 = 0$

because $X^T X z = 0 \Rightarrow X X^T X z = 0 \Rightarrow Xz = 0$ since XX^T is full rank.

Moore Penrose inverse.

The solution to the minimization problem for $d \leq n$ and $d > n$ is (with $\Sigma = 0$ here)

$$\hat{\beta}_{LS} = X^{+} Y$$

where $X^{+} = \begin{cases} (X^T X)^{-1} X^T Y, & \text{if } X^T X \text{ is full rank } (d \leq n) \\ X^T (X X^T)^{-1} Y, & \text{if } X X^T \text{ is full rank } (d > n). \end{cases}$

is the Moore-Penrose inverse.

Note: in exercises we introduce the Moore-Penrose inverse in full generality. It is always defined even without assuming that $X^T X$ or $X X^T$ is full rank. However then the formula for X^{+} is different. Not it can always be obtained from the singular value decomposition.

To conclude we give another approach to the main result here.

Theorem. Least squares minimizer.

(a) If X^TX is full rank $L(\beta) = \|Y - X\beta\|^2$ has unique minimizer $\hat{\beta}_{LS} = (X^TX)^{-1}X^TY$.

(b) If XX^T is full rank $L(\beta) = \|Y - X\beta\|^2$ has minimizers $\hat{\beta}_{LS} = X^T(XX^T)^{-1}Y + z$ where $z \in$ Null space of (X^TX)

Proof

(a) X^TX full rank is invertible. So $\hat{\beta}_{LS}$ is well defined

$$\|Y - X\beta\|^2 = \|Y - X(X^TX)^{-1}X^TY + X(X^TX)^{-1}X^TY - X\beta\|^2$$

$$= \|Y - X(X^TX)^{-1}X^TY\|^2 + \|X(X^TX)^{-1}X^TY - X\beta\|^2$$

$$+ 2\langle Y - X(X^TX)^{-1}X^TY,\ X(X^TX)^{-1}X^TY - X\beta \rangle$$

Now $Y - X(X^TX)^{-1}X^TY = (I - X(X^TX)^{-1}X^T)Y$.

= vector orthogonal to row space of X^T.

since $X(X^TX)^{-1}X^T$ is a projection onto row space of X^T (see previous remarks).

and $X(X^TX)^{-1}X^T - X\beta$ belongs to row space of X^T.

Thus $\langle \, , \, \rangle = 0$.

finally $\|Y - X\beta\|^2 \geq \|(I - X(X^TX)^{-1}X^T)Y\|^2$

with equality for $\beta = \hat{\beta}_{LS}$.

(b) Proof is even simpler. Note that if

$z \in$ Nul space (X^TX) we have $X^TXz = 0$

so $XX^TXz = 0$ so $Xz = 0$ since XX^T is assumed to be full rank.

$L(\beta) = \|Y - X\beta\|^2 \geq 0$ with equality for $\beta_{LS} = X^T(XX^T)^{-1}Y + z$

just seen by replacing inside. ∎