Linear models

Ref: Bishop's book

Max-likelihood & least-squares

$(*) \rightarrow y = \underset{\uparrow \text{unknown}}{w}^T x + \epsilon$, where $\epsilon \sim N(0, \underset{\uparrow \text{precision parameter}}{\frac{1}{\beta}})$

Suppose $\mathcal{D}_n = \{ (x_i, y_i), i = 1 \ldots n \}$ iid & satisfying $(*)$.

$$\mathcal{L}(w, \beta) = \prod_{i=1}^{n} \frac{\sqrt{\beta}}{\sqrt{2\pi}} \exp\left(-\frac{\beta}{2} (y_i - w^T x_i)^2 \right)$$

$$l(w,\beta) = \log \mathcal{L}(w, \beta) = \frac{n}{2} \log \beta - \frac{n}{2} \log 2\pi - \beta \left[\frac{1}{2} \sum_{i=1}^{n} (y_i - w^T x_i)^2 \right]$$

To maximize $l(w, \beta)$ wrt w, it is enough to

$$\hat{w}_{ML} = \arg\min_{w} \frac{1}{2} \sum_{i=1}^{n} (y_i - w^T x_i)^2 \leftarrow \text{empirical risk minimization}$$

Linear Regression:

Given data $\{ (x_i, y_i); i = 1 \ldots n \}$

$x_i \in \mathbb{R}^d$, $y \in \mathbb{R}$

$$J(w) = \frac{1}{2} \sum_{i=1}^{n} (x_i^T w - y_i)^2$$

Minimize J:

$$A = \begin{bmatrix} - & x_1^T & - \\ - & x_2^T & - \\ & \vdots & \\ - & x_n^T & - \end{bmatrix}_{n \times d}$$

$$Aw = \begin{bmatrix} x_1^T w \\ \vdots \\ x_n^T w \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$Aw - Y = \begin{bmatrix} x_1^T w - y_1 \\ \vdots \\ x_n^T w - y_n \end{bmatrix}$$

$$(Aw-Y)^T (Aw-Y) = \sum_{i=1}^{n} (x_i^T w - y_i)^2 = 2J(w)$$

$$\nabla_w J(w) = \frac{1}{2} \nabla_w (Aw-Y)^T (Aw-Y) = A^T (Aw-Y)$$

So, $\nabla J(w) = 0 \iff (A^T A) w = A^T Y$

Can we write $w = (A^T A)^{-1} A^T Y$?

Yes, if A is full rank.

since full rank $A \Rightarrow A^T A$ is invertible

(Why? argue $\text{rank}(A) = \text{rank}(A^T A)$ by showing $N(A) = N(A^T A)$)

Augmented features: $\{(\tilde{x}_i, y_i), i = 1 \ldots n\}$ $\tilde{x}_i \in \mathbb{R}^d$

$$x_i = (1, \underleftrightarrow{\ \tilde{x}_i\ })$$

$$f(x) = \sum_{i=1}^{d} w_i x_i + w_0 = \underbrace{w^T \tilde{x}}_{d\text{-dimensional objects}} + w_0$$

Why w_0?

$$J(w) = \frac{1}{2} \sum_{i=1}^{n} \left(w^T \tilde{x}_i + w_0 - y_i \right)^2$$

Take partial derivative wrt w_0

$$\frac{\partial J}{\partial w_0} = 0$$

$$\sum_{i=1}^{n} \left(w^T \tilde{x}_i + w_0 - y_i \right) = 0$$

Simplifying, $w_0 = \frac{1}{n} \sum_{i=1}^{n} y_i - w^T \left(\sum_{i=1}^{n} \tilde{x}_i \right)$

<u>Polynomial regression</u>:

$d=1$, $\{ (x_i, y_i), i=1 \ldots n \}$ $x_i, y_i \in \mathbb{R}$

Use transformed features, i.e.,

the following model

$$\hat{y}(x) = \underbrace{w_0 + w_1 x + w_2 x^2 + \cdots + w_m x^m}_{\text{mth degree polynomial}}$$

features ← polynomial basis.

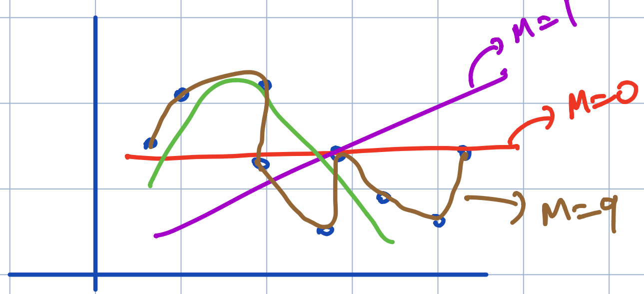
$$\hat{y}(x) = \sum_{j=0}^{m} w_j \phi_j(x), \quad \phi_j(x) = x^j$$

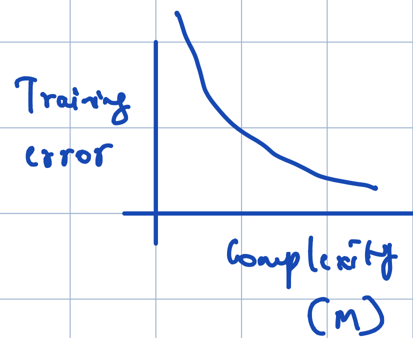
$$= w^T \phi(x)$$

$$w = (w_0, \ldots, w_m)$$

$$\phi(x) = (\phi_0(x), \ldots, \phi_m(x))$$
$$= (1, x, x^2, \ldots, x^m)$$

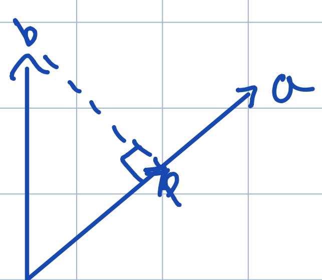
Alternately, $\phi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{2s^2}\right)$ ← Gaussian basis





Least-squares: Geometric viewpoint

Recap of projections:-



$$p = \hat{x} a$$

$$(b - \hat{x} a) \perp a$$

$$a^T(b - \hat{x}a) = 0$$

$$\hat{x} = \frac{a^T b}{a^T a}$$

$$p = \hat{x}a = \left(\frac{a^T b}{a^T a}\right) a$$

$$= a\,\frac{a^T b}{a^T a} = \left[\frac{1}{a^T a}(aa^T)\right] b$$

Projection matrix

$$P = \frac{1}{a^T a} aa^T$$

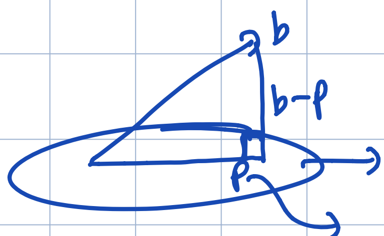
Projection matrix is (i) symmetric

(ii) $P^2 = P$

<u>Project onto a subspace:</u>

A is a $m \times n$-matrix.

Want: Project b onto $Col(A)$



$Col(A) =$ span (columns of A)

$p = A\hat{x}$

$e = b - p = b - A\hat{x}$

e $\perp$ every vector in $C(A)$

$\Rightarrow \quad e \in N(A^T)$ because $N(A^T) \perp C(A)$

$$A^T(b - A\hat{x}) = 0$$

$$A^T A\hat{x} = A^T b$$

observe $A = \begin{bmatrix} | & & | \\ a_1 & - - - & a_n \\ | & & | \end{bmatrix}$

$$a_1^T(b - A\hat{x}) = 0$$
$$\vdots$$
$$a_n^T(b - A\hat{x}) = 0$$

$$\begin{bmatrix} a_1^T \\ \vdots \\ a_n^T \end{bmatrix} \begin{bmatrix} b - A\hat{x} \end{bmatrix} = 0$$

$$A^T(b - A\hat{x}) = 0$$

So, $\quad A^T A\hat{x} = A^T b$ & this minimizes

$$E = \|Ax - b\|^2$$

If A is full col-rank, then

$$\hat{x} = (A^T A)^{-1} A^T b$$

Projection $\quad p = A\hat{x} = A(A^T A)^{-1} A^T b$

Example:-

$$\begin{bmatrix} -1 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} \theta' \\ \theta'' \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 3 \end{bmatrix}$$

Want to solve $\quad A\theta = b$

Check if b is in $C(A)$.

$$\left[\begin{array}{cc|c} -1 & 1 & 1 \\ +1 & 1 & 1 \\ 2 & 1 & 3 \end{array}\right] \longrightarrow \cdots\cdots \longrightarrow \left[\begin{array}{cc|c} 1 & -1 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{array}\right]$$

inconsistent system

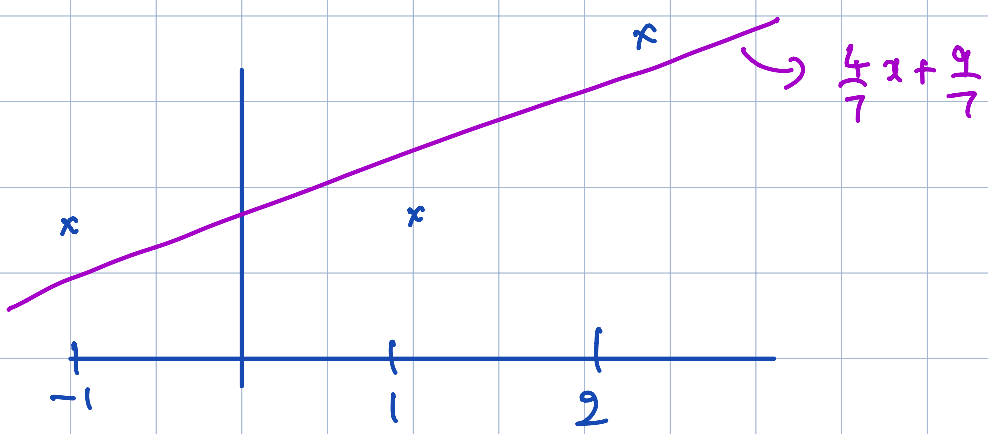
So $b \notin C(A)$.

Least-squares: $\quad A^T A \, \hat{\theta} = A^T b$

$$A^T A = \begin{bmatrix} 6 & 2 \\ 2 & 3 \end{bmatrix}$$

$$\begin{bmatrix} 6 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} \hat{\theta}' \\ \hat{\theta}'' \end{bmatrix} = \begin{bmatrix} 6 \\ 5 \end{bmatrix}$$

$$\hat{\theta}' = \frac{4}{7} \qquad \hat{\theta}'' = \frac{9}{7}$$



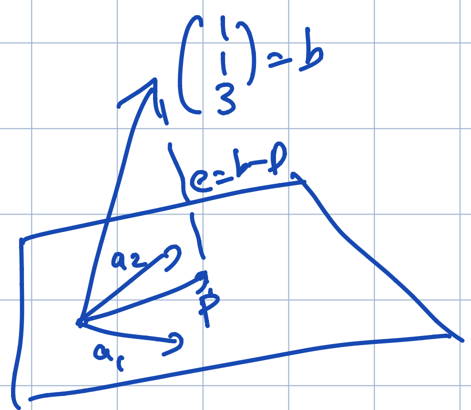
$$P_1 = \frac{5}{7}, \quad P_2 = \frac{13}{7}, \quad P_3 = \frac{17}{7}$$

$$E^2 = \|b - A\hat{\theta}\|^2$$

$$= \left[1 - \left(-\frac{4}{7} + \frac{9}{7}\right)\right]^2 + \left[1 - \left(\frac{4}{7} + \frac{9}{7}\right)\right]^2 + \left[3 - \left(\frac{8}{7} + \frac{9}{7}\right)\right]^2$$

$$= +\frac{2}{7}$$

$$e = b - p = \left(\frac{2}{7}, -\frac{6}{7}, \frac{4}{7}\right)$$



$$e \perp \underset{a_1}{\underset{\shortparallel}{(-1, 1, 2)}} \qquad e \perp \underset{a_2}{\underset{\shortparallel}{(1, 1, 1)}}$$

BIAS-VARIANCE TRADEOFF

$f(\cdot) \rightarrow$ predictor i.e., $f(x)$ is the prediction for input feature x, & y is the target

Suppose (x, y) are chosen using some distribution $\mathcal{D}$.

Then, the expected loss or the "risk" is

$$R(f) = \underbrace{E}_{\text{over } x, y}\left((f(x) - y)^2 \right)$$

Claim: $f^*(x) = E[y|x]$ is the best predictor,

ie., f^* minimizes $R(f)$.

Pf: $E\left((f(x) - y)^2 \right)$

$$= E\left[\left(f(x) - E[y|x] + E[y|x] - y \right)^2 \right]$$

$$(*) \quad E\left((f(x) - y)^2 \right) = E\left(f(x) - E[y|x] \right)^2 + E\left[\left(E(y|x) - y \right)^2 \right]$$

Predictor f present only in this term

noise term

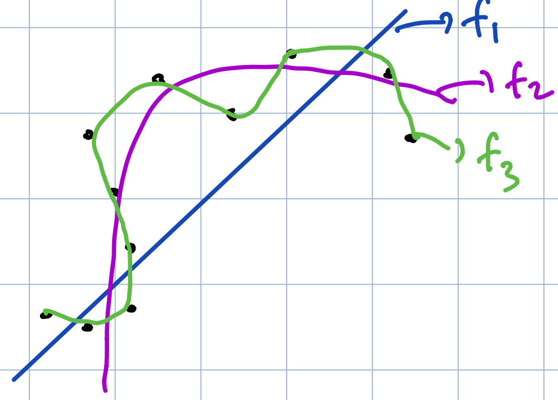
$(*)$ is minimized for $f = E[y|x]$ ▨

In a typical ML setting, $R(f)$ cannot be evaluated for a given f, since the underlying distributions are unknown.

So, collect training data $\{ (x_i, y_i), i = 1 \ldots n \}$

sampled iid from $\mathcal{D}$, and minimize

Empirical risk $\rightarrow$ $$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \left(f(x_i) - y_i \right)^2$$



Intuitively,

f_1 is very simple

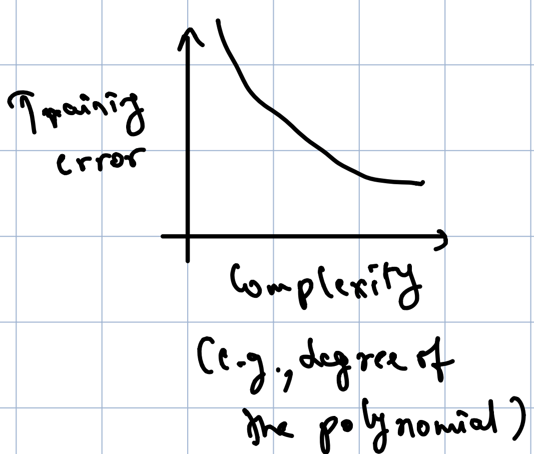
f_3 is very accurate

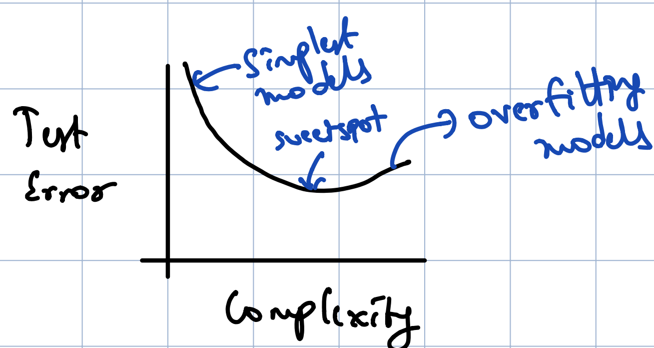
Maybe f_2 is the right fit

So, it is not enough to use $R_n(f)$ to judge f, since f_3 minimizes $R_n(f)$ (& fits noise). This phenomenon is referred to as "overfitting".

$$\text{Test (or generalization) error for } f = \frac{1}{m}\sum_{i=1}^{m}\left(\bar{y}_i - f(\bar{x}_i)\right)^2, \text{ where}$$

$\{(\bar{x}_i, \bar{y}_i), i = 1 \dots m\}$ is the test data generated using distribution $\mathcal{D}$ (used for generating training data as well).





PTO

Goal: $\min_{f} E\left((f_D(x)-y)^2\right)$

$f_D(\cdot) \rightarrow$ predictor learnt using some dataset D.

$$E\left((f_D(x)-y)^2\right)$$

$$= \underbrace{E\left((f_D(x)-E[y|x])^2\right)}_{(I)} + \underbrace{E\left((E(y|x)-y)^2\right)}_{\substack{\text{noise term} \\ \text{(unavoidable)}}}$$

$$(I) = E\left((f_D(x)-E(y|x))^2\right)$$

$$= E\Big[(f_D(x)-E_D(f_D(x)))^2 + (E_D(f_D(x))-E(y|x))^2$$
$$+ 2(f_D(x)-E_D(f_D(x)))(E_D(f_D(x))-E(y|x))\Big]$$

$E_D(f_D(x)) = E(f_D(x)|x)$

Fix x & average over many datasets, say $D_1, D_2, \ldots$

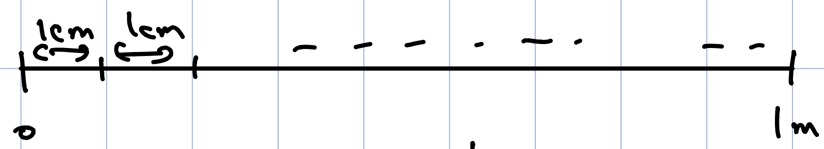
$$= E\left[(f_D(x)-E_D(f_D(x)))^2\right] \rightsquigarrow \text{Variance}$$

$$+ E\left[\left[E_D(f_D(x))-E(y|x)\right]^2\right] \rightsquigarrow (\text{Bias})^2$$

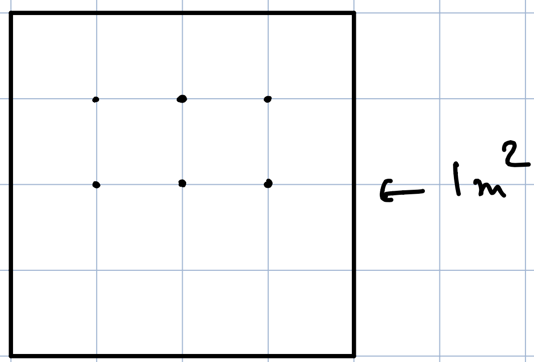
It is easy to see that as f grows complex, the bias decreases.

Claim: As f grows complex, variance increases.

"Curse of dimensionality".



100 points



10^k points

points required to sample a unit hypercube grows exponentially with the dimension

$\min_{f \in \mathcal{F}_1} \hat{R}(f)$, $\min_{f \in \mathcal{F}_2} \hat{R}(f)$, -- & so on, where

$\mathcal{F}_i$ = set of all polynomials with degree at most i.

Vector of Co-efficients of a polynomial in $\mathcal{F}_i$ sit in $\mathbb{R}^{d+1}$

With increasing i, one needs to explore more points to find the best f in $\mathcal{F}_i$

(OR)

Given a fixed # of points, the parameter space is explored less efficiently for higher order $\mathcal{F}_i$, leading to errors.

A workaround: Add a penalty cost, i.e., solve the following variant of ERM:

$$\frac{1}{2}\sum_{j=1}^{n}(y_j - f(x_j))^2 + \lambda \underbrace{L(i)}_{\text{complexity cost}} \quad \text{---} (*)$$

From ML discussion before,

$$\min \frac{1}{2}\sum_{j=1}^{n}(y_j - f(x_j))^2 \iff \max \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{(y_i - f(x_i))^2}{2}\right)$$

where $y_i = f(x_i) + \epsilon_i$

$\epsilon_i \hookrightarrow$ Standard Gaussian

Similarly, $(*)$ can be viewed as

$$\min \frac{1}{2}\sum_{j=1}^{n}(y_j - f(x_j))^2 + \lambda L(i)$$

$$\Updownarrow$$

$$\max \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y_j - f(x_j))^2}{2}\right) \times \underbrace{C(\lambda)\exp(-\lambda L(i))}_{\text{prior probability on } \mathcal{F}_i}$$

A choice for $L(i)$: $L(i) = \|\beta(i)\|^2$

Regularized version of regression:- (Ridge regression)

$$\hat{R}_n(f) = \min_{w} \frac{1}{2} \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|^2$$

Too small a $\lambda \to$ no effect of regularization (overfit)

Too large a $\lambda \to$ under fit

Ref: Table 1.2

<u>Illustration of bias-variance tradeoff for a linear model:</u>

Consider the model $y = w^T x + \epsilon$, $\epsilon \sim N(0, \frac{1}{\beta})$

The ML estimate $\hat{w}_{ML}$ for w, given $\mathcal{D} = \{(x_i, y_i), i=1 \ldots n\}$

$$\hat{w}_{ML} = (A^T A)^{-1} A^T Y, \text{ where}$$

$$A = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \eta = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

"Assume cols of A are linearly independent".

(a) $E(\hat{w}_{ML}) = E\left((A^T A)^{-1} A^T y \right)$

$= (A^T A)^{-1} A^T E y$

$= (A^T A)^{-1} A^T A w$

$= w.$

(b) $Var(\hat{w}_{ML})$

$= E\left((\hat{w}_{ML} - w)(\hat{w}_{ML} - w)^T \right)$

$= E\left[\left((A^T A)^{-1} A^T y - w \right) \left((A^T A)^{-1} A^T y - w \right)^T \right]$

$= E\left[\left((A^T A)^{-1} A^T y - w \right) \left(y^T A (A^T A)^{-1} - w^T \right) \right]$

$= (A^T A)^{-1} A^T E[y y^T] A (A^T A)^{-1} - w w^T$

$= (A^T A)^{-1} A^T E\left[(Aw + \eta)(Aw + \eta)^T \right] A (A^T A)^{-1}$
$- w w^T$

$= (A^T A)^{-1} A^T \left[A w w^T A^T + \frac{1}{\beta} I_{n \times n} \right] A (A^T A)^{-1}$
$- w w^T$

$= w w^T + \frac{1}{\beta} (A^T A)^{-1} - w w^T$

$= \frac{1}{\beta} (A^T A)^{-1}$

(c) Bias-variance decomposition

$$E\left((f_D(x) - y)^2 \right)$$

$$= E_{x,y}\left(\underbrace{(E(y|x) - y)^2}_{\text{noise}} \right) + E_x\left[\left(\underbrace{E_D(f_D(x)) - E(y|x)}_{\text{bias}} \right)^2 \right]$$

$$+ E\left(\underbrace{f_D(x) - E_D(f_D(x))}_{\text{Variance}})^2 \right)$$

for linear case, $f_D(x) = \hat{w}_{ML}^T x$

$$\text{Bias} = E_D(\hat{w}_{ML}^T x) - w^T x = 0 \quad \left(\begin{array}{c}\text{from}\\ \text{part (a)}\end{array}\right)$$

$$\text{Variance} = E\left((f_D(x) - E_D(f_D(x)))^2 \,\middle|\, x \right)$$

Let $\tilde{E} = E(\cdot|x)$

$$= \tilde{E}\left((x^T \hat{w}_{ML} - x^T w)^2 \right)$$

$$= \tilde{E}\left((x^T (A^T A)^{-1} A^T Y - x^T w)^2 \right)$$

$$= \tilde{E}\left((x^T (A^T A)^{-1} A^T (Aw + \eta) - x^T w)^2 \right)$$

$$= \tilde{E}\left((x^T w + x^T (A^T A)^{-1} A^T \eta - x^T w)^2 \right)$$

$$= \tilde{E}\left(\left(x^T (A^TA)^{-1} A^T \eta \right)^2 \right)$$

$$= \tilde{E}\left(\left(x^T (A^TA)^{-1} A^T \eta\right)\left(x^T (A^TA)^{-1} A^T \eta\right)^T \right)$$

$$= x^T (A^TA)^{-1} A^T \underbrace{E\left(\eta\eta^T\right)}_{= \frac{1}{\beta} I_{n \times n}} \left(x^T (A^TA)^{-1} A^T\right)^T$$

$$= \frac{1}{\beta} x^T (A^TA)^{-1} A^T A (A^TA)^{-1} x$$

$$= \frac{1}{\beta} x^T (A^TA)^{-1} x$$

H.W.

① Let $C = A(A^TA)^{-1}A^T$

Show that C is a projection matrix, i.e., C is symmetric & $C^2 = C$

② Check if $(I - C)$ is a projection matrix

③ Let $S = \text{span}(\text{cols of } A)$. Show that, for any $z \in \mathbb{R}^d$, Cz is the projection of z onto S.

④ Show that $A\hat{w}_{ML}$ is orthogonal to $Y - A\hat{w}_{ML}$.

H.W. Redo the exercise for a ridge regression-based estimate $\hat{w}_{Reg}$.

(i) Write out the expression for $\hat{w}_{Reg}$

(i) Write out the expression for $\hat{w}_{Reg}$

given $\mathcal{D} = \{(x_i, y_i), i = 1 - n\}$

(ii) $E(\hat{w}_{Reg})$, $Var(\hat{w}_{Reg})$

(iii) Calculate the bias & variance components.