1) H is symmetric (i.e., $H^T = H$).

$$H^T = (X(X^TX)^{-1}X^T)^T$$
$$= X((X^TX)^{-1})^T X^T \qquad [\because (ABC)^T = C^TB^TA^T]$$
$$= X((X^TX)^T)^{-1} X^T \qquad [\because (X^{-1})^T = (X^T)^{-1}]$$
$$= X(X^TX)^{-1}X^T$$
$$= H.$$

b) $H^K = H$

$$H^2 = X(X^TX)^{-1}X^T X(X^TX)^{-1}X^T$$
$$= X(X^TX)^{-1}\underbrace{(X^TX)(X^TX)^{-1}}_{I}X^T$$
$$= X(X^TX)^{-1}$$

for $K=1$, $H^1 = H$

Suppose, $H^K = H$

Then $H^{K+1} = H^K \cdot H$

$$= H\cdot H = X(X^TX)^{-1}X^T X(X^TX)^{-1}X^T$$
$$= X(X^TX)^{-1}\underbrace{(X^TX)(X^TX)^{-1}}_{I}X^T$$
$$= X(X^TX)^{-1}X^T = H.$$

} principle of mathematical induction

c) $(I-H)^K = (I-H)$.

when $K=1 \Rightarrow (I-H)^1 = I-H$

Suppose, $(I-H)^K = (I-H)$.

Then $(I-H)^{K+1} = (I-H)^K(I-H)$

$$= (I-H)(I-H)$$
$$= I - 2H + H^2$$
$$= I - 2H + H \qquad (\text{from 1b} \because H^2 = H).$$
$$= I - H.$$

} principle of mathematical induction.

d) trace$(H) = d+1$.

$$\text{trace}(\underbrace{X}_{A}\underbrace{(X^TX)^{-1}X^T}_{B}) = \text{trace}((X^TX)^{-1}X^TX) \qquad (\because tr(AB) = tr(BA))$$
$$= \text{trace}(I_{(d+1)\times(d+1)}) = d+1.$$

2a) $w_{lin} = (X^TX)^{-1}X^T\vec{y}$

$= (X^TX)^{-1}X^T(Xw^* + \vec{\epsilon})$

$\hat{\vec{y}} = Xw_{lin} = X(X^TX)^{-1}X^T(Xw^* + \vec{\epsilon})$

$= Xw^* + H\vec{\epsilon}$

alternatively,

$H = X(X^TX)^{-1}X^T$

$\hat{\vec{y}} = H\vec{y}$

$= X(X^TX)^{-1}X^T(Xw^* + \vec{\epsilon})$

$= Xw^* + H\vec{\epsilon}$.

2

Note : bold y $= \vec{y}$, $X_{N\times(d+1)}$

bold ϵ $= \vec{\epsilon}$. $w_{(d+1)\times 1}$

b) $\hat{\vec{y}} - \vec{y} = Xw^* + H\vec{\epsilon} - (Xw^* + \vec{\epsilon})$

$= H\vec{\epsilon} - \vec{\epsilon} = (H-I)\vec{\epsilon}$.

c) $E_{in}(w_{lin}) = \frac{1}{N}\|\hat{\vec{y}} - \vec{y}\|^2$

$= \frac{1}{N}(\hat{\vec{y}} - \vec{y})^t(\hat{\vec{y}} - \vec{y})$

$= \frac{1}{N}\left[(H-I)\vec{\epsilon}\right]^t\left[(H-I)\vec{\epsilon}\right]$

$= \frac{1}{N}\vec{\epsilon}^t(H-I)^t(H-I)\vec{\epsilon}$

$= \frac{1}{N}\vec{\epsilon}^t(I-H)^2\vec{\epsilon}$ ~~(from 1C)~~

$= \frac{1}{N}\vec{\epsilon}^t(I-H)\vec{\epsilon}$. (from 1c)

d) $E_D\left[E_{in}(w_{lin})\right] = E_D\frac{1}{N}\left[\vec{\epsilon}^t(I-H)\vec{\epsilon}\right]$

$= \frac{1}{N}E_D\left[\vec{\epsilon}^t\vec{\epsilon} - \vec{\epsilon}^tH\vec{\epsilon}\right]$

$\cancel{= \frac{1}{N}E_D\left[\sum_{i=1}^{N}\epsilon_i^2 - \sum_{i=1}\right]}$

$E_D[\epsilon^t\epsilon] = E_D\left[\sum_{i=1}^{N}\epsilon_i^2\right] = \sum E_D[\epsilon_i^2]$

$= N\sigma^2$

$\because E[\epsilon_i^2] = var(\epsilon_i) + (E(\epsilon_i))^2$

$E[\epsilon_i^2] = \sigma^2 + 0 = \sigma^2$

③

$$E_D[\vec{\epsilon}^t H \vec{\epsilon}] = E_D\left[\sum \epsilon_i H_{ij} \epsilon_j\right]$$

$$= \sum_{i,j=1}^{N} E_D[\epsilon_i H_{ij} \epsilon_j] \quad (\because \text{expectation \& finite sum commute})$$

$$= \sum_{i,j=1}^{N} H_{ij} E_D[\epsilon_i \epsilon_j] \quad (H_{ij} \text{ does not depend on } \epsilon_i, \epsilon_j)$$

$$= \sum_{\substack{i,j=1 \\ i=j}}^{N} h_{ii} E_D[\epsilon_i^2] + \underbrace{\sum_{i\neq j} h_{ij} E_D[\epsilon_i \epsilon_j]}_{= E_D[\epsilon_i] E_D[\epsilon_j] = 0.}$$

(as ϵ_i & ϵ_j are independent)

$$= \sum_{i=1}^{N} h_{ii} \sigma^2 = (d+1)\sigma^2 \quad (\because tr(H) = d+1)$$

combining, we get

$$\frac{1}{N}(N\sigma^2 - (d+1)\sigma^2) = \sigma^2\left(1 - \frac{d+1}{N}\right).$$

e) $E_{test}(w_{lin}) = \frac{1}{N}\|\hat{\vec{y}} - \vec{y}\|^2$

$$= \frac{1}{N}\|Xw^* + H\vec{\epsilon} - (Xw^* - \vec{\epsilon}')\|$$

$$= \frac{1}{N}\|H\vec{\epsilon} - \vec{\epsilon}'\|$$

$$= \frac{1}{N}(H\vec{\epsilon} - \vec{\epsilon}')^t(H\vec{\epsilon} - \vec{\epsilon}')$$

$$\cancel{= \frac{1}{N}(\vec{\epsilon}'^t - H}$$

$$= \frac{1}{N}(\vec{\epsilon}^t H^t - \vec{\epsilon}'^t)(H\vec{\epsilon} - \vec{\epsilon}').$$

$$= \frac{1}{N}\vec{\epsilon}^t H^t H\vec{\epsilon} - \vec{\epsilon}^t H^t \vec{\epsilon}' - \vec{\epsilon}'^t H\vec{\epsilon} + \vec{\epsilon}'^t \vec{\epsilon}'$$

$$= \frac{1}{N}(\vec{\epsilon} H\vec{\epsilon} - 2\vec{\epsilon}'^t H\vec{\epsilon} + \vec{\epsilon}'^t \vec{\epsilon}'). \quad \left[\because H^2 = H,\ H^t = H\right)$$

similar to (d),

$$E(\vec{\epsilon}'^t \vec{\epsilon}') = N\sigma^2$$

$$E(\vec{\epsilon} H \vec{\epsilon}) = (d+1)\sigma^2.$$

Now, $E_{D,\vec{\epsilon}'}\left(\vec{\epsilon}'^{t} H \vec{\epsilon}\right)$.

$$= E_D\left[E_{\vec{\epsilon}'|D}\left(\vec{\epsilon}'^{t} H \vec{\epsilon}\right)\right]$$

$$= E_D\left[E_{\vec{\epsilon}'|D}\left(\vec{\epsilon}'^{t}\right) H \vec{\epsilon}\right] \quad \left[\because H\vec{\epsilon} \text{ is independent of } \vec{\epsilon}'\right]$$

$$= E_D\left[0 \cdot H \vec{\epsilon}\right]. \quad \left(\text{as } E[\epsilon_i'] = 0\right) \quad \left(\Rightarrow E_{\vec{\epsilon}'|D}\left(\vec{\epsilon}'^{t}\right) = \vec{0}'\right)$$

$$= 0$$

Combining,

$$N\sigma^2 + (d+1)\sigma^2 = \frac{1}{N}\left(N\sigma^2 + (d+1)\sigma^2\right)$$

$$= \sigma^2\left(1 + \frac{d+1}{N}\right).$$

Question 3

a) $\epsilon \rightarrow$ noise realization of test point

$\vec{\epsilon} \rightarrow$ bold epsilon i.e., noise realization of all train data points

$$y = w^{*T}x + \epsilon$$
$$= x^T w^* + \epsilon$$

$$w_{lin} = (X^T X)^{-1} X^T y_{data}$$
$$= (X^T X)^{-1} X^T (X w^* + \vec{\epsilon})$$
$$= w^* + (X^T X)^{-1} X^T \vec{\epsilon}.$$

$$\Rightarrow g(x) = x^t\left(w^* + (X^T X)^{-1} X^T \vec{\epsilon}\right).$$
$$= x^t w^* + x^t (X^T X)^{-1} X^T \vec{\epsilon}$$

$$\Rightarrow y - g(x) = \epsilon - x^t (X^T X)^{-1} X^T \vec{\epsilon}.$$

b) Take joint expectation w.r.t x & ϵ.

$$E_{out} = E_{X,\epsilon}\left[(y-g(x))\right]^2$$

$$= E_{X,\epsilon}\left[(y-g(x))^t(y-g(x))\right]$$

$$= E_{X,\epsilon}\left[\left(\epsilon - x^t(X^TX)^{-1}X^T\vec{\epsilon}\right)^t\left(\epsilon - x^t(X^TX)^{-1}X^T\vec{\epsilon}\right)\right].$$

$$= E_{X,\epsilon}\left[\left(\epsilon - \vec{\epsilon}^{\,t}X(X^TX)^{-1}x\right)\left(\epsilon - x^t(X^TX)^{-1}X^T\vec{\epsilon}\right)\right]$$

(as ϵ is scalar $\epsilon^t = \epsilon$)

$$= E_{x,\epsilon}\left[\epsilon^2 - \epsilon\, x^t(X^TX)^{-1}X^T\vec{\epsilon} - \vec{\epsilon}^{\,t}X(X^TX)^{-1}x\,\epsilon + \vec{\epsilon}^{\,t}X(X^TX)^{-1}x x^t(X^TX)^{-1}X^T\vec{\epsilon}\right]$$

$$= E_{X,\epsilon}[\epsilon^2] - E_{x,\epsilon}\left[2\epsilon\, x^t(X^TX)^{-1}X^T\vec{\epsilon}\right] + E_{x,\epsilon}\left[\vec{\epsilon}^{\,t}X(X^TX)^{-1}x x^t(X^TX)^{-1}X^T\vec{\epsilon}\right]$$

(as 2nd & 3rd terms are scalar & for scalar a, $a = a^t$)

1st term ~~$E_{x,\epsilon}$~~

$$E_{x,\epsilon}[\epsilon^2] = E_x\, E_{\epsilon|x}[\epsilon^2] = E_x[\sigma^2] = \sigma^2$$

2nd term

$$E_{x,\epsilon}\left[2\epsilon\, x^t(X^TX)^{-1}X^T\vec{\epsilon}\right] = 2\, E_x\left[E_{\epsilon|x}\, \epsilon\, \underbrace{x^t(X^TX)^{-1}X^T\vec{\epsilon}}_{\text{independent of } \epsilon \,\&\, x}\right]$$

$$= 2\, E_x\left[\cancel{E_{\epsilon|x}[\epsilon]}^{\,0}\ \epsilon\, x^t(X^TX)^{-1}X^T\vec{\epsilon}\right]$$

$$= 0.$$

3rd term:-

$$E_{x,\epsilon}\left[\vec{\epsilon}^{\,t}X(X^TX)^{-1}x x^t(X^TX)^{-1}X^T\vec{\epsilon}\right]$$

$$= E_{x,\epsilon}\left[tr\left(\underbrace{\vec{\epsilon}^{\,t}X(X^TX)^{-1}x}_{A}\ \underbrace{x^t(X^TX)^{-1}X^T\vec{\epsilon}}_{B}\right)\right]$$

(as the term inside trace is scalar)

$$= E_{x,\epsilon}\left\{ tr\left[x x^t (X^TX)^{-1} X^T \vec{\epsilon}\vec{\epsilon}^t X (X^TX)^{-1} \right]\right\}$$ [$tr(AB) = tr(BA)$] 6

$$= tr\ E_{x,\epsilon}\left[\quad \text{"} \quad \right]$$ [as expectation & trace commute]

$$= tr\left\{ E_x\left[E_{\epsilon|x}(x x^t)\right] \underbrace{(X^TX)^{-1} X^T \vec{\epsilon}\vec{\epsilon}^t X (X^TX)^{-1}}_{\text{independent of both } \epsilon \,\&\, x.} \right\}$$

$$= tr\left[E_x(x x^t) \left[(X^TX)^{-1} X^T \vec{\epsilon}\vec{\epsilon}^t X (X^TX)^{-1}\right]\right]$$

$$= tr\left[\Sigma (X^TX)^{-1} X^T \vec{\epsilon}\vec{\epsilon}^t X (X^TX)^{-1}\right]$$

c) $E_{\vec{\epsilon}}\left[\vec{\epsilon}\vec{\epsilon}^t\right] = E_{\vec{\epsilon}}\begin{bmatrix} \epsilon_1^2 & \epsilon_1\epsilon_2 & \cdots & \epsilon_1\epsilon_n \\ \epsilon_2\epsilon_1 & \epsilon_2^2 & \cdots & \epsilon_2\epsilon_n \\ \vdots & & & \\ \epsilon_n\epsilon_1 & & \cdots & \epsilon_n\epsilon_n \end{bmatrix}$

$$E_{\vec{\epsilon}}\left[\epsilon_i^2\right] = \sigma^2$$

$$E_{\vec{\epsilon}}\left[\epsilon_i\epsilon_j\right] = E_{\epsilon_i}\left[\cancel{\epsilon_i}\right]^0 E_{\epsilon_j}\left[\cancel{\epsilon_j}\right]^0 = 0.$$

~~So, $\sigma^2 I_{N\times N}$~~

as ϵ is of dimension $N \times 1$.

$$E_{\vec{\epsilon}}\left[\vec{\epsilon}\vec{\epsilon}^t\right] = \sigma^2 I_{N\times N}.$$

as $\vec{\epsilon}$ is of ~~dimension~~ $N \times 1$, ~~the resulting matrix~~ we have $\cancel{E_{\vec{\epsilon}}}\left[\vec{\epsilon}\,\vec{\epsilon}^{\,T}\right] = \sigma^2 I_{N\times N}$

(3) (7)

d) From (c),

$$E_{out} = \sigma^2 + \text{trace}\left[\Sigma (X^TX)^{-1} X^T \vec{\epsilon}\,\vec{\epsilon}^{\,T} X (X^TX)^{-1}\right]$$

$$\Rightarrow E_{\vec{\epsilon}}[E_{out}] = \sigma^2 + E_{\vec{\epsilon}}\ \text{tr}\left[\qquad\qquad\right]$$

$$= \sigma^2 + \text{tr}\ E_{\vec{\epsilon}}\left[\qquad\qquad\right]$$

$$= \sigma^2 + \text{tr}\left(\underbrace{\Sigma (X^TX)^{-1} X^T}_{P}\ E_{\vec{\epsilon}}\left[\vec{\epsilon}\,\vec{\epsilon}^{\,T}\right]\ \underbrace{X (X^TX)^{-1}}_{Q}\right].$$

P & Q are independent of $\vec{\epsilon}$.

$$\cancel{= \sigma^2 + \text{tr}\left[\Sigma (X^TX)^{-1} X^T X (X^TX)^{-1}\right]}$$

$$= \sigma^2 + \text{tr}\left[\Sigma (X^TX)^{-1} X^T \sigma^2 I\ X (X^TX)^{-1}\right]$$

$$= \sigma^2 + \sigma^2\ \text{tr}\left[\Sigma (X^TX)^{-1} (X^TX) (X^TX)^{-1}\right]$$

$$= \sigma^2 + \sigma^2\ \text{tr}\left(\Sigma (X^TX)^{-1}\right)$$

$$= \sigma^2 + \frac{\sigma^2}{N}\ \text{tr}\left(\Sigma \left(\frac{X^TX}{N}\right)^{-1}\right).$$

If $\frac{X^TX}{N} = \Sigma$ then

$$= \sigma^2 + \frac{\sigma^2}{N}\ \text{tr}\left(\Sigma \Sigma^{-1}\right)$$

$$= \sigma^2 + \frac{\sigma^2}{N}(d+1) = \sigma^2\left(1 + \frac{d+1}{N}\right).$$

e) X^TX is an N-Sample estimate of Σ.

By the law of large numbers $\frac{X^TX}{N} \xrightarrow{p} \Sigma$.

using continuity at Σ, $\left(\frac{X^TX}{N}\right)^{-1} \xrightarrow{p} \Sigma^{-1}$

$$\Rightarrow \Sigma\left(\frac{X^TX}{N}\right)^{-1} \xrightarrow{p} \Sigma\Sigma^{-1} = I$$

$$\Rightarrow \text{trace}\left(\Sigma\left(\frac{X^TX}{N}\right)^{-1}\right) \xrightarrow{p} d+1$$ ($\because$ as trace is continuous at I)

$$\Rightarrow \text{trace}\left(\Sigma\left(\frac{X^TX}{N}\right)^{-1}\right) = (d+1) + \epsilon.$$ for som small scalar ϵ.

$$\Rightarrow \sigma^2 + \frac{\sigma^2}{N}\,\text{trace}\left(\Sigma\left(\frac{1}{N}X^TX\right)^{-1}\right) = \sigma^2\left(1 + \frac{d+1}{N} + \frac{\epsilon}{N}\right).$$

$$\Rightarrow E_{out} = \sigma^2\left(1 + \frac{d+1}{N} + o\left(\frac{1}{N}\right)\right).$$

COL341 Spring 2023
Homework 1
Grading guidelines

1 Question1

- Everyone has got good marks in this.
- Question 1a, mention the result $((AB)^{-1})^T = ((AB)^T)^{-1}$ i.e., transpose and inverse commute.
- One can formalize the proof in 1b and 1c by using mathematical induction properly.

2 Question2

- You should make the distinction between ϵ (noise associated with the test input) and $\boldsymbol{\epsilon}$ (noise realizations of the training examples)
- In Q2d, mention the independence of ϵ_i and ϵ_j, $\mathbb{E}[\epsilon] = 0$ and $\mathbb{E}[\epsilon^2] = var(\epsilon) - [\mathbb{E}[\epsilon]]^2 = \sigma^2$, where you are using these results.
- Mention the independence of ϵ and ϵ'.
- Mention which random variable are you taking the expectation over, and break the join distribution appropriately, In 2e, $\mathbb{E}_{\mathcal{D},\epsilon'}$ should be broken into $\mathbb{E}_{\mathcal{D}}[\mathbb{E}_{\epsilon'|\mathcal{D}}]$.

3 Question3

-
- You should make the distinction between ϵ (noise associated with the test input) and $\boldsymbol{\epsilon}$ (noise realizations of the training examples)
- In 3c, mention teh dimension of the matrix.
- Mention in which step did you use, $a = trace(a)$ for any scalar and $trace(AB) = trace(BA)$ and what is your A and B.

- Mention trace and expectation commute.
- Some have mentioned that test noise ϵ is independent of test point x, which is not true.
- In 3b, $\mathbb{E}_{\boldsymbol{x},\epsilon'}$ should be broken into $\mathbb{E}_{\boldsymbol{x}}[\mathbb{E}_{\epsilon'|\boldsymbol{x}}]$.
- Some have not done the last bit of question 3d and 0.25 have been deducted for that.
- Additionally, mention how can you take certain terms out of the expectation.
- In 3e, mention the continuity of trace at the identity matrix $\mathbb{I}$.