Assignment 8

Name: Mohaiminul Al Nahian
RIN: 662026703
Course no.: CSCI 6100

Exercise 4.3

a) Assuming H is fixed and we increase the complexity of f, then the deterministic noise will in general go up. As a result, there will be a higher tendency to overfit.

b) If f is fixed and we decrease the complexity of H, then the deterministic noise will go up in general. But simpler model will have a huge gain in generalization error. This will result in better out of sample error, even though the in-sample error might be bigger as compared to complex H. Because decreasing the complexity will make it less susceptible to fitting noise. So, there will be a lower tendency to overfit.

Exercise 4.5

a) IF $\Gamma = I^{Q+1}$, where I^{Q+1} is an identity matrix with dimension $Q+1$, we obtain

$$w^T \Gamma^T \Gamma w = w^T \mathrm{I}^T \mathrm{I} w = w^T w = \sum_{q=0}^{Q} w_q^2 \leq C$$

b) If $\Gamma = [1, 1, ..., 1]$ is a Q+1 dimensional vector of ones, we can obtain:

$$\begin{aligned} w^T\Gamma^T\Gamma w &= [w_0 \dots w_Q] \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} [1 \dots 1] \begin{bmatrix} w_0 \\ \vdots \\ w_Q \end{bmatrix} \\ &= (w_0 + w_1 + \cdots + w_Q)(w_0 + w_1 + \cdots + w_Q) \\ &= \sum_{q=0}^{Q} w_q \sum_{q=0}^{Q} w_q = \left(\sum_{q=0}^{Q} w_q \right)^2 \leq C \end{aligned}$$

Exercise 4.6

The hard-order constraint should be more useful in terms of binary classification. Since, $sign(w^T x) = sign(\alpha w^T x)$ for all $\alpha > 0$, so both αw and w will draw same separator line/plane. So, limiting the length of vector will help the model to choose simpler solution. But as hard order constraint will restrict the model to use some of its weight, there will be less tendency to overfit. So, using hard-order constraint would be better in this case.

Exercise 4.7

a)

$$\begin{aligned} \sigma_{\text{val}}^2 &= \text{Var}_{\mathcal{D}_{\text{val}}} \left[E_{\text{val}}(g^-) \right] \\ &= \text{Var}_{\mathcal{D}_{\text{val}}} \left[\frac{1}{K} \sum_{x_n \in \mathcal{D}_{\text{val}}} e(g^-(x_n), y_n) \right] \\ &= \frac{1}{K^2} \text{Var}_{\mathcal{D}_{\text{val}}} \left[\sum_{x_n \in \mathcal{D}_{\text{val}}} e(g^-(x_n), y_n) \right] \\ &= \frac{1}{K^2} \left[\sum_{x_n \in \mathcal{D}_{\text{val}}} Var_x[e(g^-(x_n), y_n)] \right] \\ &= \frac{1}{K^2} \cdot K \cdot \text{Var}_{\text{x}} \left[e(g^-(x), y) \right] \\ &= \frac{1}{K^2} \cdot K \cdot \sigma^2(g^-) \\ &= \frac{1}{K} \sigma^2(g^-) [Showed] \end{aligned}$$

b) Given, $e(g^{-}(x), y) = [g^{-}(x) \neq y]$
Now,

$$P\left[e(g^{-}(x), y) = 1\right] = P\left[g^{-}(x) \neq y\right] = p$$
$$P\left[e(g^{-}(x), y) = 0\right] = P\left[g^{-}(x) = y\right] = 1 - p$$

So, expected value of $e(g^{-}(x), y)$ is $E[e(g^{-}(x), y)] = 1 * p + 0 * (1 - p) = p$.
So,

$$\begin{aligned}\sigma^2_{\text{val}} = \text{Var}_{\mathcal{D}_{\text{val}}}\left[E_{\text{val}}(g^{-})\right] &= \text{Var}_{\mathcal{D}_{\text{val}}}\left[\frac{1}{K}\sum_{x_n \in \mathcal{D}_{\text{val}}} e(g^{-}(x_n), y_n)\right] \\ &= \frac{1}{K} \cdot \text{Var}_{\text{x}}\left[e(g^{-}(x), y)\right] \\ &= \frac{1}{K}(E[e^2] - (E[e])^2)\end{aligned}$$

We know that $e = e^2$, So,

$$\begin{aligned}\sigma^2_{\text{val}} &= \frac{1}{K}(E[e] - (E[e])^2) \\ &= \frac{p - p^2}{K}[showed]\end{aligned}$$

c)

$$\begin{aligned}\sigma^2_{\text{val}} &= \frac{p - p^2}{K} \\ &= \frac{1}{K}\left[\frac{1}{4} - (p - \frac{1}{2})^2\right]\end{aligned}$$

So, when $p = \frac{1}{2}$ we get the maximum value of σ^2_{val}, which is is $\dfrac{1}{4K}$, thus $\sigma^2_{val} \leq \dfrac{1}{4K}$ [showed]

d) Here, $e(g^{-}(x), y) = (g^{-}(x) - y)^2$
And, there will be no upper bound on $\text{Var}_{\mathcal{D}_{\text{val}}}\left[E_{\text{val}}(g^{-})\right]$
We can express it as in (b)

$$\text{Var}_{\mathcal{D}_{\text{val}}}\left[E_{\text{val}}(g^{-})\right] = \frac{1}{K}(E[e^2] - (E[e])^2)$$

Since the squared error e is unbounded, in this case this variance is composed of unbounded terms. So, there is no theoretical uniform upper bound for $Var\left[E_{val}(g^{-})\right]$

e) If we train using fewer data, the $\sigma^2(g^{-})$ is expected to be higher. Since fewer data points indicates our g^{-} will be a bad approximation to target function and so expected error (or, mean) should be higher. And from the hint, higher 'mean' often indicates higher variance. So, variance is expected to increase.

f) Since $\sigma^2_{\text{val}} = \dfrac{1}{K}\sigma^2(g^{-})$, increasing the size of validation set means increasing K. so, the variance should decrease and so do E_{out}. But increasing K decreases the number of training sample. So, we may come up with a worse g that increases $\sigma^2(g^{-})$. Thus increasing the size of the validation set can result in both a better or a worse estimate of E_{out}, depending on the dominating effect of the two factors stated above.

Exercise 4.8

Yes, E_m is an unbiased estimate for the out of sample error $E_{out}(g^{-}_m)$, because validation set is not involved in the training process.

Problem 4.26

a) We have, $z_n = \phi(x_n)$ has a dimention of $d \times 1$. So, $z_n z_n^T$ has dimention $d \times d$. Let, $z_n = \begin{bmatrix} z_{n1} \\ z_{n2} \\ \vdots \\ z_{nd} \end{bmatrix}$ and,

Let, Z be the following matrix containing z_k^T [where, k=1,2,....,N] in each row,

$$Z = \begin{bmatrix} z_1^T \\ z_2^T \\ \vdots \\ z_N^T \end{bmatrix} \text{ so, } Z^T = [z_1, z_2, \ldots z_N]$$

$$So, Z^T Z = [z_1, z_2, \ldots, z_N] \begin{bmatrix} z_1^T \\ z_2^T \\ \vdots \\ z_N^T \end{bmatrix} = \sum_{n=1}^{N} z_n z_n^T [Showed] \tag{1}$$

$$and, Z^T y = [z_1, z_2, \ldots z_N] \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \sum_{n=1}^{N} z_n y_n [Showed] \tag{2}$$

$$Now, H(\lambda) = ZA(\lambda)^{-1} Z^T = Z(Z^T Z + \lambda \Gamma^T \Gamma)^{-1} Z^T = \begin{bmatrix} z_1^T \\ z_2^T \\ \vdots \\ z_N^T \end{bmatrix} (Z^T Z + \lambda \Gamma^T \Gamma)^{-1} [z_1, z_2, \tag{3}$$

$$= \begin{bmatrix} z_1^T \\ z_2^T \\ \vdots \\ z_N^T \end{bmatrix} \left[A(\lambda)^{-1} z_1, A(\lambda)^{-1} z_2, \ldots \right. \tag{4}$$

From the above equation it is evident that

$$H_{nm}(\lambda) = z_n^T (Z^T Z + \lambda \Gamma^T \Gamma)^{-1} z_m = z_n^T A(\lambda)^{-1} z_m [Showed]$$

And when (z_n, y_n) is left out, from equation (1) and (2) we can write that $Z^T Z \rightarrow Z^T Z - z_n z_n^T$ and $Z^T y \rightarrow Z^T y - z_n y_n$

b) From the hint, $(A - xx^T)^{-1} = A^{-1} + \dfrac{A^{-1} x x^T A^{-1}}{1 - x^T A^{-1} x}$, we replace x by z_n. We shall replace x with z.
We know that,

$$w = (Z^T Z + \lambda \Gamma^T \Gamma)^{-1} Z^T y = A^{-1} Z^T y$$

If w_n^- is the weight vector when the nth data point is left out, we can write,

$$w_n^- = A_{-n}^{-1} Z_{-n}^T y_{-n}$$

Here, $A_{-n}^{-1} = (A - z_n z_n^T)^{-1}$ and $Z_{-n}^T y_{-n} = Z^T y - z_n y_n$. So,

$$\begin{aligned}
w_n^- &= (A - z_n z_n^T)^{-1}(Z^T y - z_n y_n) \\
&= (A^{-1} + \frac{A^{-1} z_n z_n^T A^{-1}}{1 - z_n^T A^{-1} z_n})(Z^T y - z_n y_n)[Showed]
\end{aligned}$$

c) From (b),

$$\begin{aligned}
w_n^- &= (A^{-1} + \frac{A^{-1} z_n z_n^T A^{-1}}{1 - z_n^T A^{-1} z_n})(Z^T y - z_n y_n) \\
&= A^{-1} Z^T y + \frac{A^{-1} z_n z_n^T A^{-1}}{1 - z_n^T A^{-1} z_n} Z^T y - A^{-1} z_n y_n - \frac{A^{-1} z_n z_n^T A^{-1}}{1 - z_n^T A^{-1} z_n} z_n y_n \\
&= w + \frac{A^{-1} z_n z_n^T w}{1 - H_{nn}} - A^{-1} z_n y_n - \frac{A^{-1} z_n H_{nn} y_n}{1 - H_{nn}}[From(a)] \\
&= w + \frac{A^{-1} z_n \hat{y_n}}{1 - H_{nn}} - \frac{A^{-1} z_n y_n - A^{-1} z_n H_{nn} y_n + A^{-1} z_n H_{nn} y_n}{1 - H_{nn}} \\
&= w + \frac{A^{-1} z_n \hat{y_n}}{1 - H_{nn}} - \frac{A^{-1} z_n y_n}{1 - H_{nn}} \\
&= w + \frac{\hat{y_n} - y_n}{1 - H_{nn}} A^{-1} z_n [Showed]
\end{aligned}$$

d) From (c),

$$\begin{aligned}
w_n^- &= w + \frac{y_n - \hat{y_n}}{1 - H_{nn}} A^{-1} z_n \\
or, z_n^T w_n^- &= z_n^T(w + \frac{y_n - \hat{y_n}}{1 - H_{nn}} A^{-1} z_n) \\
&= \hat{y_n} + \frac{\hat{y_n} - y_n}{1 - H_{nn}} z_n^T A^{-1} z_n \\
&= \hat{y_n} + \frac{\hat{y_n} - y_n}{1 - H_{nn}} H_{nn} \\
&= \frac{\hat{y_n} - \hat{y_n} H_{nn} + \hat{y_n} H_{nn} - y_n H_{nn}}{1 - H_{nn}} \\
&= \frac{\hat{y_n} - H_{nn} y_n}{1 - H_{nn}} [Showed]
\end{aligned}$$

e)

$$\begin{aligned}
e_n &= (z_n^T w_n^- - y_n)^2 \\
&= (\frac{\hat{y_n} - H_{nn} y_n}{1 - H_{nn}} - y_n)^2 \\
&= (\frac{\hat{y_n} - H_{nn} y_n - y_n + H_{nn} y_n}{1 - H_{nn}})^2 \\
&= (\frac{\hat{y_n} - y_n}{1 - H_{nn}})^2 [Showed]
\end{aligned}$$

We know that cross validation estimate is the average value of e_n's

$$\begin{aligned}
Or, E_{cv} &= \frac{1}{N} \sum_{n=1}^{N} e_n \\
&= \frac{1}{N} \sum_{n=1}^{N} (\frac{\hat{y_n} - y_n}{1 - H_{nn}})^2
\end{aligned}$$

Which proves equation 4.13