CSCI 4100 Assignment 8

Boliang Yang 661541863

1. Exercise 4.3 in LFD
   (a) Since the deterministic noise is the part of target function outside the best fit, then increasing the complexity of f will increase the part that H cannot model. Thus, the deterministic noise will go up.

   Since complexity of f increases and H will fit more deterministic noise, then there is a higher tendency to overfit.
   (b) If we decrease the complexity of H, then it will increase the part that H cannot model. Thus, the deterministic noise will go up.

   Since f is fixed and the complexity of H decreases, then H will be simpler comparing to f. Thus, there is lower tendency to overfit.

2. Exercise 4.5 in LFD
   (a) Since $\sum_{q=0}^{Q} w_q^2 = w^T w = w^T I^T I w \leq C$ and $w^T \Gamma^T \Gamma w \leq C$, then we have $\Gamma = I$, which is the identity matrix.
   (b) Since $w^T \Gamma^T \Gamma w = (\Gamma w)^T (\Gamma w) \leq C$ and $\left(\sum_{q=0}^{Q} w_q\right)^2 \leq C$, then we can choose any matrix Γ with one row of ones, $\Gamma = [1\ 1\ \cdots\ 1\ 1]$. Then, $\Gamma w = [1\ 1\ \cdots\ 1\ 1]\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_{Q-1} \\ w_Q \end{bmatrix} = \sum_{q=0}^{Q} w_q$, and it follows that $w^T \Gamma^T \Gamma w = \left(\sum_{q=0}^{Q} w_q\right)^2 \leq C$.

3. Exercise 4.6 in LFD

   I expect the hard-order constraint to be more useful for binary classification using the perceptron model. For classification, hard-order constraint will decrease the higher order parameters which lead to a significant decrease in VC dimension. But since for any $\alpha > 0$, $sign(w^T x) = sign(\alpha w^T x)$, then soft-order constraint cannot influence the result of the classification. Therefore, the VC dimension does not have a significant decrease in soft-order constraint.

4. Exercise 4.7 in LFD

(a) $\sigma_{val}^2 \stackrel{\text{def}}{=} Var_{D_{val}}[E_{val}(g^-)] = Var_{D_{val}}\left[\frac{1}{K}\sum_{x_n \in D_{val}} e(g^-(x_n), y_n)\right] =$

$\frac{1}{K} Var_{D_{val}}\left[\sum_{x_n \in D_{val}} e(g^-(x_n), y_n)\right] = \frac{1}{K} Var_x[e(g^-(x), y)] = \frac{1}{K}\sigma^2(g^-)$

(b) Since $e(g^-(x), y) = [\![g^-(x) \neq y]\!] = \begin{cases} 0 \ if \ g^-(x) = y \\ 1 \ if \ g^-(x) \neq y \end{cases}$, then $E[e(g^-(x), y)] =$

$P[g^-(x) \neq y]$, $E[e^2(g^-(x), y)] = P[g^-(x) \neq y]$. Then we can express σ_{val}^2 as:
$\sigma_{val}^2 = \frac{1}{K} Var_x[e(g^-(x), y)] = \frac{1}{K}(E[e^2(g^-(x), y)] - E[e(g^-(x), y)]^2) =$
$\frac{1}{K}(P[g^-(x) \neq y] - P[g^-(x) \neq y]^2)$

(c) Since from part (b) we have, $\sigma_{val}^2 = \frac{1}{K}(P[g^-(x) \neq y] - P[g^-(x) \neq y]^2) =$

$\frac{1}{K}\left(\frac{1}{4} - \left(P[g^-(x) \neq y] - \frac{1}{2}\right)^2\right)$, and also $P[g^-(x) \neq y] \in [0,1]$, then it follows

that $\sigma_{val}^2 = \frac{1}{K}\left(\frac{1}{4} - \left(P[g^-(x) \neq y] - \frac{1}{2}\right)^2\right) \in \left[0, \frac{1}{4K}\right]$. Thus, $\sigma_{val}^2 \leq \frac{1}{4K}$.

(d) No. Since the square error is unbounded, then there is no upper bound for $E[e(g^-(x), y)]$. Thus, there is no upper bound for $Var_x[e(g^-(x), y)]$ either.

(e) I expect $\sigma^2(g^-)$ to be higher. Since for all x, $e(g^-(x), y) = (g^-(x) - y)^2 \geq 0$, then if we train using fewer points, $g^-(x)$ will create more errors and $E[e(g^-(x), y)]$ will increase. Also, since for continuous, non-negative random variables, higher mean often implies higher variance, then it follows that the variance will increase.

(f) Worse. Since increasing the size of the validation set will decrease the size of the training set, then it will result in a worse estimate of E_{out}.

5. Exercise 4.8 in LFD

Yes. By the definition of the validation set, we know D_{val} will not influence the actual training. Also, since $E_{D_{val}}[E_{val}(g_m^-)] = E_{out}(g_m^-)$, then it is an unbiased estimate.