CSCI-4100 Assignment 8
Yichuan Wang
RIN:661414395

EXERCISES
4.3
(a) For a fixed H with hypotheses that are less complex than target function f, if target function becomes more complex, the deterministic noise goes up. Since our model cannot distinguish noises, it has a higher tendency to overfit.

(b) For a fixed target function, if we assume that the complexity of our model is less than that of the target, then decreasing the complexity of H will increase the deterministic noise and thus increase the tendency to overfit. On the other hand, decreasing the model complexity will make it less likely to fit noise since there are simply less ways to go wrong, and therefore lower the tendency to overfit.
As shown in Figure 4.2 in the text book, a simpler model with a higher deterministic noise can have a huge gain from its robustness against noise when there is less data, and a more complex model will gain from its lower deterministic noise when there is more data to battle against noises. So the tendency to overfit for a simpler model depends on how much data we have; with more data the tendency will increase since more deterministic noise plays the major role, with less data the tendency will decrease since the robustness against noise plays the major role.

4.5
(a) $\Gamma = I^{d+1}$ would suffice, where I^{d+1} is an identity matrix with $d+1$ dimension. In this case we have

$$w^T\Gamma^T\Gamma w = w^T w \leq C$$

which by definition is equivalent to $\sum_{q=0}^{Q} w_q^2 \leq C$

(b) For this problem Γ can be a $(d+1) \times (d+1)$ matrix with all 1's on its first row and all other elements being 0. We have $w^T\Gamma^T\Gamma w = (w^T\Gamma^T)(\Gamma w)$ and with our Γ we can derive that $(w^T\Gamma^T)$ becomes a row vector with its first element being $\sum_{q=0}^{Q} w_q$ and other elements being 0, and (Γw) becomes a column vector with its first element being $\sum_{q=0}^{Q} w_q$ and other elements being

0. As a result we have

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

4.6 Hard-order constraint should be more useful. We know that w and αw draws the same line/hyperplane in the space as long as $\alpha > 0$. Limiting the length of vector w won't do anything on limiting the model to choose a complex solution since any weight vector that points to the same direction is the same separator; the model is still able to draw all possible separators regardless of whether there is a soft-order constraint. On the other hand, a hard-order constraint can forbid the model to use some of its weights, and thus limiting its tendency to overfit.

4.7
(a)

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

(b)

$$\sigma^2_{val} = \frac{1}{K}Var_x[e(g^-(x_n), y_n)]$$
$$= \frac{1}{K}(E[e(g^-(x_n), y_n)^2] - E[e(g^-(x_n), y_n)]^2)$$

Since we define $e(g^-(x), y) = [\![g^-(x) \neq y]\!]$, we know that $e = e^2$ according to the definition of inversion bracket $[\![]\!]$, and thus we can derive

$$\sigma_{val}^2 = \frac{1}{K}(E[e(g^-(x_n), y_n)] - E[e(g^-(x_n), y_n)]^2)$$

We know by definition that $E[e(g^-(x_n), y_n)] = P[e(g^-(x_n), y_n)] = P[g^-(x) \neq y]$, so we have

$$\frac{1}{K}(E[e(g^-(x_n), y_n)] - E[e(g^-(x_n), y_n)]^2) = \frac{1}{K}(P[g^-(x) \neq y] - P[g^-(x) \neq y]^2)$$

(c) We can use the result from b:

$$\sigma_{val}^2 = \frac{1}{K}(P - P^2)$$

$$= \frac{1}{K}(-P^2 + P - \frac{1}{4} + \frac{1}{4})$$

$$= \frac{1}{K}(\frac{1}{4} - (P - \frac{1}{2})^2)$$

It is obvious that the maximum of this term is $\frac{1}{4K}$ with the square term being zero.

(d) No. In this case we can express $Var[E_{val}(g^-)]$ as following: (similar to calculation in part b)

$$Var[E_{val}(g^-)] = \frac{1}{K}(E[e]^2 - E[e^2])$$

We know that e is the squared error and is thus unbounded, and thus we know that this variance is composed of unequal and unbounded terms. In this case there is no theoretical upper bound for this variance.

(e) Higher. With less training points our g^- will be a worse approximation to target function, indicating a higher expected error (the mean), and according to hint, higher mean often implies higher variance.

(f)We have $\sigma_{val}^2 = \frac{1}{K}\sigma^2(g^-)$. When K is small, increasing the size of validation set will decrease the variance dramatically due to the fact that K is

in the denominator and the expected error won't increase much since $(N-K)$ is still relatively large, and thus E_{out} will be lower overall. When K is large, the fact that K is in the denominator won't help much, and the fact that the training set is small makes the increment in error mean and error variance play the major role in E_{out}, so E_{out} will be larger.

4.8 Yes, E_m is an unbiased estimate since the validation set is not involved in the training of any individual model.

PROBLEMS

4.26
(a)
$Z^TZ = \sum_{n=1}^{N} z_n z_n^T$: According to matrix multiplication, the jth element of ith row of Z^TZ is the sum of all $z_n[i]z_n[j]$, and since the jth element of ith row of $z_n^T z_n$ is $z_n[i]z_n[j]$ for this particular z_n, Z^TZ represents the sum of all $z_n z_n^T$, which is $\sum_{n=1}^{N} z_n z_n^T$.
$Z^Ty = \sum_{n=1}^{N} z_n y_n$: The ith row of Z corresponds to the ith elements of all z_n. So Z^Ty represents a column vector in which every element corresponds to the sum of corresponding elements of all z_n times y_n, which if written in math language is $\sum_{n=1}^{N} z_n y_n$
From above discussion we know that when (z_n,y_n) is left out, $Z^TZ \longrightarrow Z^TZ - z_n z_n^T$ and $Z^Ty \longrightarrow Z^Ty - z_n y_n$