Machine Learning from Data CSCI 4100
Assignment 8
Jae Park (RIN: 661994900)

Exercise 4.3

a) If the complexity of f goes up while $\mathcal{H}$ remains fixed, then the deterministic noise will go up because f gets more complex relative to $\mathcal{H}$. Because the complexity goes up and noise increases, this will increase the tendency to overfit.

b) If the complexity of $\mathcal{H}$ decreases while f remains fixed, this will make the deterministic noise to go up because now our model is relatively simpler.
With regard to overfitting, a less complex hypothesis $\mathcal{H}$ in general will lower the tendency to overfit since it is more likely to *ignore* the noise. However, this relies on the assumption that our data size is relatively moderate. If the data is large enough, the noise from them will cancel out the effect of the simpler model and increase the tendency to overfit. So, whether a simpler model will overfit or not largely depends on the size of the data set.

Exercise 4.5

a) We know $\mathbf{w}^T\mathbf{w} \leq C$, and if we had $\Gamma = \mathrm{I}_{Q+1}$,

$$
\begin{aligned}
\mathbf{w}^T\Gamma^T\Gamma\mathbf{w} &= \mathbf{w}^T\mathrm{I}_{Q+1}^T\mathrm{I}^{Q+1}\mathbf{w} \\
&= (\mathrm{I}_{Q+1}\mathbf{w})^T\mathrm{I}_{Q+1}\mathbf{w} \\
&= \left(\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ & & & \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} \mathbf{w}_0 \\ \vdots \\ \mathbf{w}_Q \end{bmatrix} \right)^T \cdot \left(\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ & & & \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} \mathbf{w}_0 \\ \vdots \\ \mathbf{w}_Q \end{bmatrix} \right) \\
&= [\mathbf{w}_0 \dots \mathbf{w}_Q] \begin{bmatrix} \mathbf{w}_0 \\ \vdots \\ \mathbf{w}_Q \end{bmatrix} \\
&= \sum_{q=0}^{Q} \mathbf{w}_q^2 \\
\therefore \sum_{q=0}^{Q} \mathbf{w}_q^2 &\leq C
\end{aligned}
$$

b) If $\Gamma = [1, ..., 1]$,

$$\mathbf{w}^T\Gamma^T\Gamma\mathbf{w} = \left([\mathbf{w}_0 \dots \mathbf{w}_Q] \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \right) \cdot \left([1 \dots 1] \begin{bmatrix} \mathbf{w}_0 \\ \vdots \\ \mathbf{w}_Q \end{bmatrix} \right)$$
$$= \sum_{q=0}^{Q} \mathbf{w}_q \times \sum_{q=0}^{Q} \mathbf{w}_q = (\sum_{q=0}^{Q} \mathbf{w}_q)^2$$

Exercise 4.6

In some cases like optimization problems, soft order constraint can be more useful than the hard order constraint. However, for the case of binary classification, we can see that it has no effect on the model.

The hard order constraint is useful in a sense that it can require some weights in $\mathbf{w}$ to be zero and make the model less susceptible to noise. However, as long as $\alpha > 0$,
$\mathbf{w}^T$ and $\alpha\mathbf{w}^T$ produce the *equivalent* vectors in the same direction, only with different magnitudes. So, $sign(\mathbf{w}^T\mathbf{x}) = sign(\alpha\mathbf{w}^T\mathbf{x})$ always holds, and the soft constraint basically does nothing in the classification.

Therefore, the hard order constraint is expected to be more useful for binary classification using the perceptron model.

Exercise 4.7

a) Given $(X, Y) = \{(\mathbf{x}_i, y_i) | \, \mathbf{x}_i \perp\!\!\!\perp y_i\}$

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

$$\therefore \sigma^2_{\text{val}} = \frac{1}{K}\sigma^2(g^-)$$

b) Given $e(g^-(\mathbf{x}), y) = [\![g^-(\mathbf{x}) \neq y]\!]$, by the notion of Iverson bracket, we let

$$\begin{aligned}
&\mathbb{P}[g^-(\mathbf{x}) \neq y] = \mathbb{P}\left[e(g^-(x), y) = 1\right] = p \\
&\quad \text{for the case} \quad g^-(\mathbf{x}) \neq y
\end{aligned}$$

$$\begin{aligned}
&\mathbb{P}[g^-(\mathbf{x}) = y] = \mathbb{P}\left[e(g^-(x), y) = 0\right] = (1-p) \\
&\quad \text{for the case} \quad g^-(\mathbf{x}) = y
\end{aligned}$$

Then we have,

$$\begin{aligned}
\sigma^2_{\text{val}} &= \frac{1}{K^2} \cdot K \cdot \text{Var}_{\mathbf{x}}\left[e(g^-(\mathbf{x}), y)\right] \quad \text{from part (a)} \\
&= \frac{1}{K} \cdot \left[\text{E}_{\mathbf{x}}\left[e(g^-(\mathbf{x}), y)^2\right] - \text{E}_{\mathbf{x}}\left[e(g^-(\mathbf{x}), y)\right]^2\right]
\end{aligned}$$

because the function $e(g^-(\mathbf{x}), y)$ is closed under square operation, i.e., $e(g^-(\mathbf{x}), y) = (e(g^-(\mathbf{x}), y))^2$ with the same domain and its range $\in \{0, 1\}$.

$$= \frac{1}{K} \cdot \left[\text{E}_{\mathbf{x}}\left[e(g^-(\mathbf{x}), y)\right] - \text{E}_{\mathbf{x}}\left[e(g^-(\mathbf{x}), y)\right]^2\right]$$

Now, the expected value of $e(g^-(\mathbf{x}), y)$ is

$$\mathrm{E}_{\mathbf{x}}\left[e(g^-(\mathbf{x}), y)\right] = 1 \times p + 0 \times (1-p) = p$$

$$\begin{aligned} So, \quad &= \frac{1}{K} \cdot \left[\mathrm{E}_{\mathbf{x}}\left[e(g^-(\mathbf{x}), y)\right] - \mathrm{E}_{\mathbf{x}}\left[e(g^-(\mathbf{x}), y)\right]^2\right] \\ &= \frac{1}{K} \cdot \left[p - p^2\right] \\ &= \frac{p - p^2}{K} \end{aligned}$$

$$\therefore \sigma^2_{\text{val}} = \frac{\mathbb{P}\left[g^-(\mathbf{x}) \neq y\right] - \mathbb{P}\left[g^-(\mathbf{x}) \neq y\right]^2}{K}$$

c)

$$\begin{aligned} \sigma^2_{\text{val}} &= \frac{p - p^2}{K} \qquad \text{where } p = \mathbb{P}[g^-(\mathbf{x}) \neq y] \\ &= \frac{1}{K}\left[-p^2 + p\right] \\ &= \frac{1}{K}\left[-(p - \frac{1}{2})^2 + \frac{1}{4}\right] \\ &= \frac{1}{4K} - \frac{(p - \frac{1}{2})^2}{K} \end{aligned}$$

The expression is a maximum when $p = \frac{1}{2}$, so the upper-bound is given as $\sigma^2_{val} \leq \dfrac{1}{4K}$.

d)
Given $e(g^-(\mathbf{x}), y) = (g^-(\mathbf{x}) - y)^2$
From (b), we have

$$\sigma^2_{\text{val}} = \frac{1}{K} \cdot \left[\mathrm{E}_{\mathbf{x}}\left[e(g^-(\mathbf{x}), y)^2\right] - \mathrm{E}_{\mathbf{x}}\left[e(g^-(\mathbf{x}), y)\right]^2\right]$$

if we replace the first Expected value term with the given, we get

$$\mathrm{E}_{\mathbf{x}}\left[e(g^-(\mathbf{x}), y)^2\right] = \mathrm{E}_{\mathbf{x}}\left[(g^-(\mathbf{x}) - y)^4\right]$$

This is an expression of a square of the unbounded squared error, so a uniform upper bound for $\mathrm{Var}\left[E_{\text{val}}(g^-)\right]$ cannot exist.

e) With fewer data points, we can expect $E_{\text{train}}(g^-)$ to be higher with the target function not approximating as well as before with more data points. For continuous, non-negative random variables, higher mean often implies higher variance $\sigma^2(g^-)$.

f) It depends, and there is no definite answer. From the result of (a), we know that $\sigma_{\text{val}}^2 = \frac{1}{K}\sigma^2(g^-)$. If we take more data from the training set and use it for the validation set, $\sigma^2(g^-)$ will become large (worse) but $\frac{1}{K}$ gets smaller at the same time. Without specific numbers, we cannot tell which of the mean and the variance will have a stronger influence on E_{out}. Also, the effect of the relative size of K will be similar to how the inverse proportion function behaves. For instance, different values of K within a smaller range might change σ_{val}^2 by a lot.

Exercise 4.8

Yes, E_m is an unbiased estimate for the out-of-sample error $E_{\text{out}}(g_m^-)$ because the validation set is independent from the estimation of g_m^-.