Exercise 4.3

a)

Initially H is less complex than the target function:

The deterministic noise will go up since we are getting a more complex target. There is a lower tendency to overfit; because the H stays unchanged; the H could be only more under-fitted to the target as the target becomes more complex.

Initially H is more complex than the target function:

The deterministic noise will go down in the beginning since the target function is becoming more equally complexed with H; it will go up again once the target function becomes more complex than H. The tendency of overfit will keep going down throughout the process.

b)

Initially H is less complex than the target function:

The deterministic noise will go up since the H will be less fit to the target function, creating more shaded area shown on figure 4.4. The tendency of overfit will go down as the H becomes less complexed.

Initially H is more complex than the target function:

The deterministic noise will go down in the beginning since the H is becoming more equally complexed with the target function; it will go up again once the H becomes less complex than the target function. The tendency of overfit will go down as the H becomes less complexed.

Exercise 4.5

a)

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

In order to gain the constraint, we need to satisfy

$w^T\Gamma^T\Gamma w = w^T w$

This means Γ has to be an identity matrix where dim(Γ) = dim(w), which is $\Gamma = I$.

Under this situation we are able to get

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

b)

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

In order to gain the constraint, we need to satisfy

$$\Gamma w = \sum_{q=0}^{Q} w_q$$

This means Γ needs to be row vector of ones; Γw will be the sum of every element in vector w.

Under this situation we are able to get

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

Exercise 4.6
Hard order constraints will be more useful.
As for soft order constraint, no matter to what extent we shrink the weights, the outcome will still be the same due to the property that
$sign(w^T x) = sign(\alpha w^T x)\ \forall\alpha > 0$.

As for hard order constraint, we force some terms of w into 0, in this case we will have the chance to change the outcome since $sign(w^T x)$ might be modified after changing some terms of w into 0. Hence, we could achieve our purpose of regularization by using hard order constraints.

Exercise 4.7

a)

We have $\sigma^{2}(g^{-}) = Var_{x}[e(g^{-}(x), y)]$

$$\sigma_{val}^{2} = Var_{x}[\frac{1}{K}\sum_{k=1}^{K} e(g^{-}(x_{k}), y_{k})]$$

$$\sigma_{val}^{2} = \frac{1}{K^{2}}Var_{x}[\sum_{k=1}^{K} e(g^{-}(x_{k}), y_{k})]$$

$$\sigma_{val}^{2} = \frac{1}{K^{2}}\sum_{k=1}^{K} Var_{x}[e(g^{-}(x_{k}), y_{k})]$$

$$\sigma_{val}^{2} = \frac{1}{K^{2}}\sum_{k=1}^{K} Var[e(g^{-})]$$

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

b)

From the question, we know that

$e(g^{-}(x), y) = 0$ if $g^{-}(x) = y$

$e(g^{-}(x), y) = 1$ if $g^{-}(x) \neq y$

We could get

$$E_{x}[g^{-}(x) = y] = P_{x}[g^{-}(x) = y] \times 0 = 0$$

$$E_{x}[g^{-}(x) \neq y] = P_{x}[g^{-}(x) \neq y] \times 1 = P[g^{-}(x) \neq y]$$

$$\sigma^{2}(g^{-}) = Var_{x}[e(g^{-}(x), y)]$$

$$\sigma^{2}(g^{-}) = E_{x}[(e(g^{-}(x), y) - E_{x}[(e(g^{-}(x), y))])^{2}]$$

$$\sigma^{2}(g^{-}) = E_{x}[(e(g^{-}(x), y) - P[g^{-}(x) \neq y])^{2}]$$

$$\sigma^{2}(g^{-}) = E_{x}[e(g^{-}(x), y)^{2} - 2e(g^{-}(x), y)P[g^{-}(x) \neq y] - P^{2}[g^{-}(x) \neq y]]$$

$$\sigma^{2}(g^{-}) = E_{x}[e(g^{-}(x), y)^{2}] - 2E_{x}[e(g^{-}(x), y)]E_{x}[P[g^{-}(x) \neq y]] + E_{x}[P^{2}[g^{-}(x) \neq y]]$$

$$\sigma^{2}(g^{-}) = P[g^{-}(x) \neq y] - 2P^{2}[g^{-}(x) \neq y] + P^{2}[g^{-}(x) \neq y$$

$$\sigma^{2}(g^{-}) = P[g^{-}(x) \neq y] - P^{2}[g^{-}(x) \neq y]]$$

$$\sigma_{val}^{2} = \frac{1}{K}(P[g^{-}(x) \neq y] - P^{2}[g^{-}(x) \neq y]])$$

c)

In order to gain the bound, we need to get the max value σ_{val}^{2} can get,

Take first derivative on σ_{val}^{2},

$$\frac{d\sigma_{val}^{2}}{P[g^{-}(x) \neq y]} = 1 - 2P[g^{-}(x) \neq y] = 0$$

Plug $P[g^{-}(x) \neq y] = \frac{1}{2}$ into σ_{val}^{2}, we get,

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

Hence, we get the max value σ_{val}^{2} can get, the bound will be

$$\sigma_{val}^{2} \leq \frac{1}{4K}$$

d)

No.

Since the squared error is unbounded, $Var_{x}[e(g^{-}(x), y)]$ will also be unbounded.

Hence, we can not get a unbound σ_{val}^{2} .

e)

If we use less data for training, we will get a worser g^{-}. This will result in higher $e(g^{-}(x), y)$ in regression. According to the hint, higher mean means more variance; in this case $Var_{x}[e(g^{-}(x), y)]$, which is σ_{val}^{2} will be higher.

f)

According to the inequality 4.9, as the E$_{val}$ becomes larger when increasing the size of validation data, the K also increases and $O(\frac{1}{\sqrt{K}})$ decreases. It is uncertain that the performance of the validation set of estimating the E$_{out}$ will increase or decrease.

Exercise 4.8

We have $E[E_{val}] = E_{out}(g^{-})$

Hence, $E[E_{m}] = E[E_{val}(g_{m}^{-})] = E_{out}(g_{m}^{-})$

It is not a biased estimate.