1.

(a)

We must normalize f to make it on the same level with $\sigma\epsilon_n$, otherwise, the noise will be meaningless.

$$\begin{aligned} E_{a,x}[f^2] &= E_x[E_{a|x}[f^2|x]] \\ &= E_x[Var_{a|x}[f] + (E_{a|x}[f])^2] \\ &= E_x[\sum_0^{Qf} L_q^2(x) Var_{a|x}[a_q] + (\sum_0^{Qf} L_q(x) E_{a|x}[a])^2] \end{aligned}$$

Since we select a from a standard normal distribution, then:

$$\begin{aligned} E_{a,x}[f^2] &= E_x[\sum_0^{Qf} L_q^2(x)] \\ &= \sum_0^{Qf} E_x[L_q^2(x)] \\ &= \sum_0^{Qf} \frac{1}{2} \int_{-1}^{1} L_q^2(x) dx \end{aligned}$$

By Problem 4.3(e) from LFD:

$$E_{a,x}[f^2] = \sum_0^{Qf} \frac{1}{2q+1}$$

should be 1

So we need to multiply $1/\sqrt{\sum_0^{Qf} \frac{1}{2q+1}}$ on f to normalize, which means the term to normalize by is $\sqrt{\sum_0^{Qf} \frac{1}{2q+1}}$.

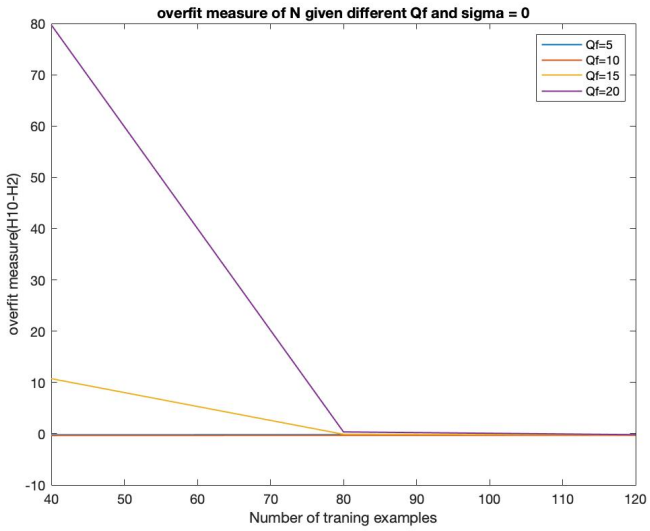
(b)

We use linear regression with each input x_n transformed to a vector V: $[L_0(x_n), L_1(x_n), ..., L_Q(x_n)]$, and each label y_n is $f(x_n) + \sigma\epsilon_n$. Then compute w_{lin} from the least square E_{in} by $w_{lin} = (X^T X)^{-1} X^T y$ where X is input matrix with each transformed vector. Finally, we use $w_{lin} V$ as g. For g_2 and g_{10}, $Q = 2$ *and* 10, respectively.

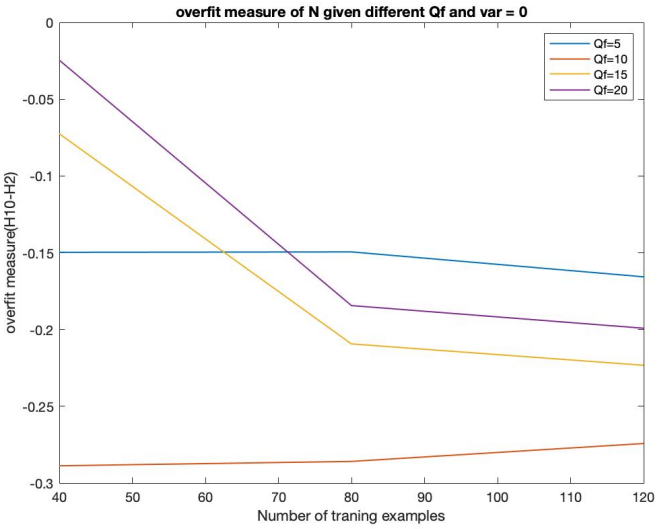
(c)

$$\begin{aligned} E_{out} &= E[(g_{10}(x) - y(x))^2] \\ &= \frac{1}{n}\sum_{i=1}^{n}(g_{10}(x_i) - y(x_i))^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(g_{10}(x_i) - f(x_i) - \sigma\epsilon_i)^2 \end{aligned}$$

(d)

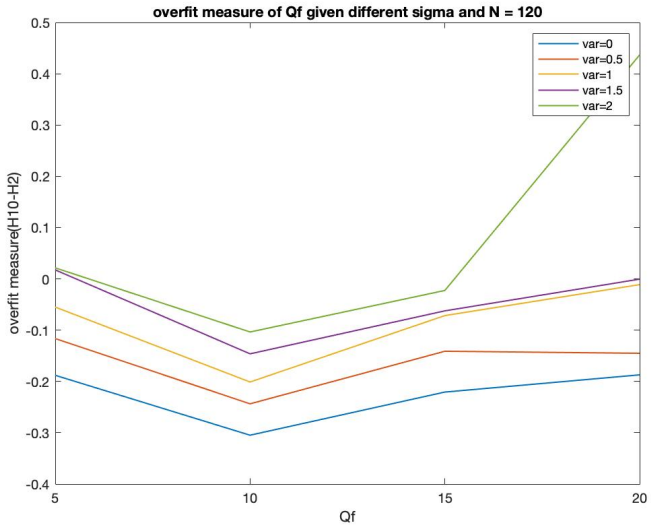


(a) Mean Eout

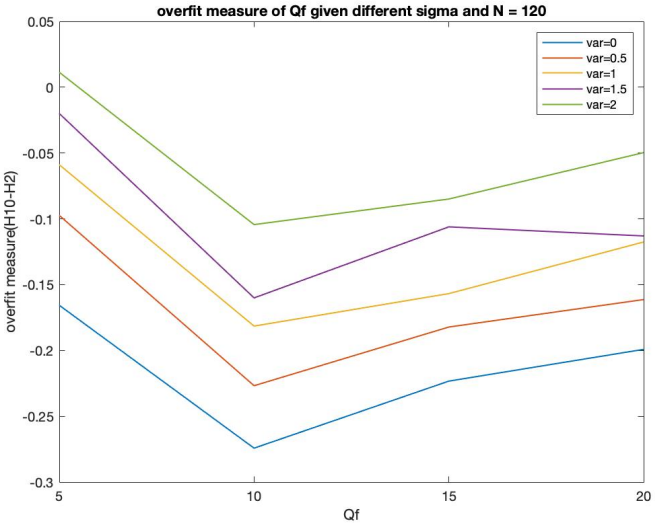


(b) Median Eout

We can see that overfit measure will decrease as training number increase given variance and Qf, and Eout of H10 will be lower than H2 if N is great enough. The high complexity of true function will make it easier overfit than low complexity of true function. Both median and mean in this case can display similar trend of overfit, but mean will be influenced by some outliers, median will lose some information on the whole data.

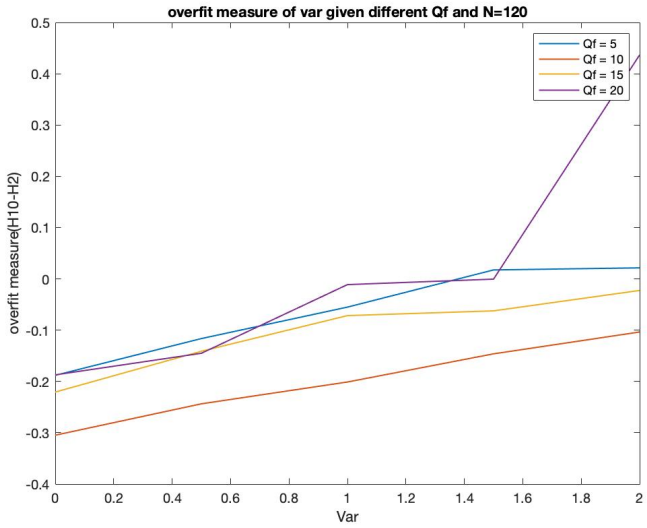


(c) Mean Eout

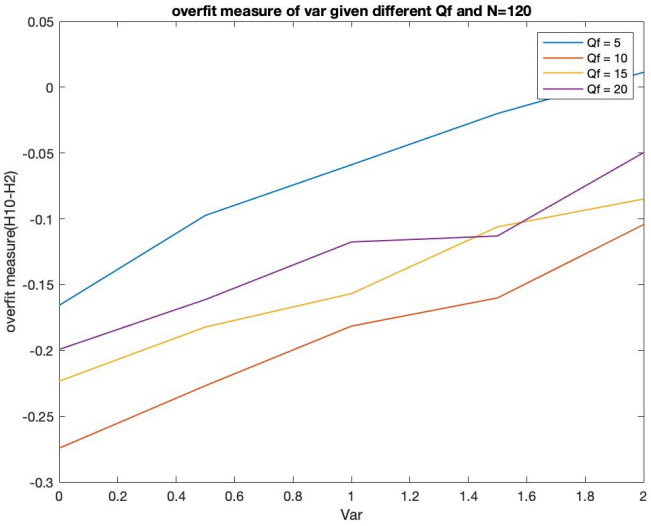


(d) Median Eout

We can see that there is trend that the more complex the true function is, the strong overfit H10 is, which means H2's generalization is better than H10. If there is not enough traning data, it will be more obvious. When variance increase, which is stochastic noise increase, the overfitting will be more severely. Moreover, we can see that when our hypothesis is close to true function's complexity level, there is less overfitting if N is large enough. Similar to above, both median and mean have similar trend, but median will lose the whole picture of data, mean will be infulenced by some outliers.



(e) Mean Eout



(f) Median Eout

We can see that if the stochastic noise increase, H10 will have more chance to overfit,, and the more complexity the true function are, the more chance to overfit. In this case, the median behaves like the zoom-in version of mean, which means it has same trend with mean, but lose the whole picture of data set, while mean acts more severely by the influence of outliers.

In conclusion, the high level of true function's complexity, high level of the stochastic noise and the high level of learning's complexity hypothesis will lead to more overfitting. While the more training data there is, the few overfitting chace it has.

2. LFD Exercise 4.5

(a)

Γ should be the identity matrix to obtain the constraint $\sum_{q=0}^{Q} w_q^2 \leq C$. Because when $\Gamma = I$, the Tikehonov regularization constraint will be $w^T w \leq C$, which is $\sum_{q=0}^{Q} w_q^2 \leq C$.

(b)

Gamma should be a vector with the same size of w full of 1. For example, if w is $R^{m\times 1}$, the Γ is $R^{1\times m}$ with all elements are 1. In this way, we can get:

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

Then:

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

3. LFD 4.8

$$E_{aug}(w) = E_{in}(w) + \lambda w^T w$$

Then:

$$\bigtriangledown E_{aug}(w(t)) = \bigtriangledown E_{in}(w(t)) + 2\lambda w(t),$$

$$\begin{aligned} w(t) - \eta \bigtriangledown E_{aug}(w(t)) &= w(t) - \eta \bigtriangledown E_{in}(w(t)) - 2\eta\lambda w(t) \\ &= (1 - 2\eta\lambda)w(t) - \eta \bigtriangledown E_{in}(w(t)) \end{aligned}$$

Proved.

4. LFD 4.25

(a)

No, each learner's validation set size is different, we can't select the minimum validation error one based on that VC-bond. (the $O(\sqrt{\frac{lnM}{2K}})$ part is different)

(b)

H_{val} is obtained before looking at the data in the validation set, the learning process is as same as learning a hypothesis from H_{val} based on D_{val}, and the $E_{val}(g_m)$ are "in-sample" errors, so we can use the VC bound for finite hypothesis. Based on the VC bound, the $O(\sqrt{\frac{lnM}{2K}})$ term is fixed when the size of validation sets are same, so we may think if there is lower $E_{val}(g_m)$, then the lower $E_{out}(g_m)$.

(c)

$$P[E_{out}(m^*) > E_{val}(m^*) + \epsilon] = P[E_{out}(m^*) - E_{val}(m^*) > \epsilon]$$

In this case, we know that $E_{out}(m^*) \geq E_{val}(m^*)$, then based on Hoeffding Inequality:

$$P[E_{out}(m^*) - E_{val}(m^*) > \epsilon] \leq e^{-2\epsilon^2 K_{m^*}}$$

$$\begin{aligned} P[E_{out}(m^*) - E_{val}(m^*) > \epsilon] &\leq P[E_{out}(1) - E_{val}(1) > \epsilon] \\ &\quad or\ P[E_{out}(2) - E_{val}(2) > \epsilon] \\ &\quad or...\ P[E_{out}(M) - E_{val}(M) > \epsilon] \\ &\leq \sum_{m=0}^{M} P[E_{out}(m) - E_{val}(m) > \epsilon] \\ &\leq \sum_{m=0}^{M} e^{-2\epsilon^2 K_m} \end{aligned}$$

We also have:

$$Me^{-2\epsilon^2 K(\epsilon)} = Me^{\frac{-2\epsilon^2}{-2\epsilon^2} ln(\frac{1}{M}\sum_{m=0}^{M} e^{-2\epsilon^2 K_m})}$$
$$= \sum_{m=0}^{M} e^{-2\epsilon^2 K_m}$$

Thus:

$$P[E_{out}(m^*) - E_{val}(m^*) > \epsilon] \leq Me^{-2\epsilon^2 K(\epsilon)}$$

Proved.

5. LFD 5.4

(a)

(i)

We did data snooping since this 500 stocks we selected already survived from 50,000 stocks during the 12,500 trading days. In other words, we pre-select this 500 from the whole 50,000 stocks, while the Hoeffding bound indicate we only select the best from 500, which is wrong.

(ii)

Based on the analysis above, we need to set $M = 50000$:

$$P[E_{out} - E_{val} > 0.02] \leq 2 \times 50000 \times e^{-2\times 12500\times 0.02^2}$$
$$\approx 4.5399$$

Which is meaningless.

(b)

(i)

Similarly to part(a), we did data snooping, we pre-select these 500 stocks from the whole 50,000 stocks, so we can not make conclusion based on these survived 500 stocks.

(ii)

If we use data of the whole 50,000 stocks during the 12,500 trading days, the conclusion we make will make sense. However, there are many stocks gone bankrupt and stopped trading, which means they don't have the whole data during 12,500 trading days. This is also a thing should be taken care with.