1. AML Problem 2.24 (page 75), except

Problem 2.24 Consider a simplified learning scenario. Assume that the input dimension is one. Assume that the input variable x is uniformly distributed in the interval $[-1, 1]$. The data set consists of 2 points $\{x_1, x_2\}$ and assume that the target function is $f(x) = x^2$. Thus, the full data set is $\mathcal{D} = \{(x_1, x_1^2), (x_2, x_2^2)\}$. The learning algorithm returns the line fitting these two points as g ($\mathcal{H}$ consists of functions of the form $h(x) = ax + b$). We are interested in the test performance (E_{out}) of our learning system with respect to the squared error measure, the bias and the var.

(a) Give the analytic expression for the average function $\bar{g}(x)$.

(b) Describe an experiment that you could run to determine (numerically) $\bar{g}(x)$, E_{out}, bias, and var.

(c) Run your experiment and report the results. Compare E_{out} with bias+var. Provide a plot of your $\bar{g}(x)$ and $f(x)$ (on the same plot).

(d) Compute analytically what E_{out}, bias and var should be.

>> Replace part (a) with:

(a.1) For a single given dataset, give an expression for $g^{(\mathcal{D})}(x)$ (AML notation).

(a.2) Find $\bar{g}(x)$ analytically; express your answer in simplest form.

>> For parts (b) and (c), obtain $\mathrm{E}_{\mathcal{D}}\{E_{out}\}$ by direct numerical computation, not by adding bias and var.

(a.1) $g(x) = ax + b$

$(x_1, x_1^2) \Rightarrow x_1^2 = ax_1 + b$

$(x_2, x_2^2) \Rightarrow x_2^2 = ax_2 + b$

$\Rightarrow a = \frac{(x_1^2 - x_2^2)}{x_1 - x_2} = x_1 + x_2$

$b = -x_1 x_2$

$\Rightarrow g^{(\mathcal{D})}(x) = (x_1 + x_2)x - x_1 x_2$

(a.2) $\bar{g}(x) = \mathbb{E}_{\mathcal{D}}\{g^{(\mathcal{D})}(x)\} = \mathbb{E}_{\mathcal{D}}\{(x_1 + x_2)x - x_1 x_2\}$

$= x(\mathbb{E}_{\mathcal{D}}[x_1] + \mathbb{E}_{\mathcal{D}}[x_2]) - \underbrace{\mathbb{E}_{\mathcal{D}}[x_1 x_2]}_{\mathbb{E}_{\mathcal{D}}[x_1] \cdot \mathbb{E}_{\mathcal{D}}[x_2]}$

$= 0$

We always assume that each point is drawn IID. Besides, since x_1 and x_2 are uniformly distributed in $[-1, 1]$,

$\mathbb{E}_{\mathcal{D}}[x_1] = \mathbb{E}_{\mathcal{D}}[x_2] = \frac{-1+1}{2} = 0$

(b) We randomly draw x_1 and x_2 in the interval $[-1, 1]$, and we can use these two points to decide a line $h(x) = \underbrace{(x_1 + x_2)}_{a}x - \underbrace{x_1 x_2}_{b}$. After repeat the experiment for 1000 times, we can get $\bar{g}(x) = \bar{a}x + \bar{b}$. Then, we can use these 1000 pairs $\mathcal{D}_i(x_1, x_2)$ to get a_i, b_i and 1000 points of x that we pick within $[-1, 1]$ with a fixed interval to calculate E_{out}, var and bias.

$E_{out} = \frac{1}{1000}\sum_{i=1}^{1000}(a_i x_i + b_i - x_i^2)^2$

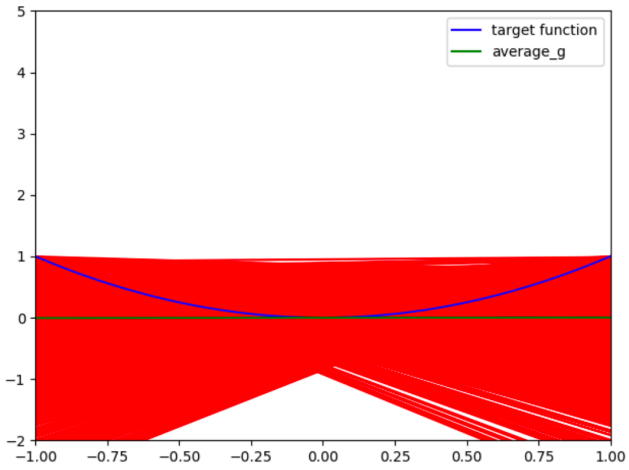
$\text{var} = \frac{1}{1000}\sum_{i=1}^{1000}\left[(a_i x_i + b_i) - (\bar{a}x_i + \bar{b})\right]^2$

$\text{bias} = \frac{1}{1000}\sum_{i=1}^{1000}\left[(\bar{a}x_i + \bar{b}) - x_i^2\right]^2$

(c)

```
g_head(x)=0.00 *x + 0.00
Eout= 0.5294279123165252
Var= 0.3333733834352551
bias= 0.20155211203794315
Var+bias= 0.5349254954731982
```

From the results above, we can find that (var+bias) is pretty close to Eout and the difference between them is approximately 0.005497.



The red part inside the plot is the lines composed from our 1000 times draw.

(d)

$$E_{out}(h_g^{\mathcal{D}}) = E_x\{[h_g^{\mathcal{D}}(x) - f(x)]^2\}$$
$$= E_x\{(ax+b-x^2)^2\}$$
$$= E_x\{x^4 + a^2x^2 - 2ax^3 - 2bx^2 + 2abx + b^2\}$$
$$= E_x\{x^4 - 2ax^3 + (a^2-2b)x^2 + 2abx + b^2\}$$
$$= \frac{1}{2}\int_{-1}^{1} x^4\,dx - 2a\cdot\frac{1}{2}\int_{-1}^{1} x^3\,dx + (a^2-2b)\cdot\frac{1}{2}\int_{-1}^{1} x^2\,dx + 2ab\cdot\frac{1}{2}\int_{-1}^{1} x\,dx + b^2$$
$$= \frac{1}{5} + \frac{a^2-2b}{3} + b^2$$

$$E_{\mathcal{D}}\{E_{out}(h_g^{\mathcal{D}})\} = E_{\mathcal{D}}\left\{\frac{1}{5} + \frac{a^2-2b}{3} + b^2\right\}$$
$$= E_{\mathcal{D}}\left\{\frac{1}{5} + \frac{x_1^2+2x_1x_2+x_2^2+2x_1x_2}{3} + x_1^2x_2^2\right\}$$
$$= \frac{1}{5} + \frac{1}{3}\left(\frac{1}{2}\int_{-1}^{1} x_1^2\,dx_1 + 2\cdot\frac{1}{4}\int_{-1}^{1} x_1+x_2\,dx_1\,dx_2 + \frac{1}{2}\int_{-1}^{1} x_2^2\,dx_2 + 2\cdot\frac{1}{4}\int_{-1}^{1}\int_{-1}^{1} x_1+x_2\,dx_1\,dx_2\right)$$
$$+ \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1} x_1^2x_2^2\,dx_1\,dx_2$$
$$= \frac{1}{5} + \frac{1}{3}\left(\frac{1}{3}+\frac{1}{3}\right) + \frac{1}{9} = \frac{8}{15}$$

$$bias = E_x\{(\underbrace{\bar{g}(x)}_{0} - f(x))^2\} = E_x\{x^4\} = \frac{1}{2}\int_{-1}^{1} x^4\,dx = \frac{1}{5}$$

$$var = E_{\mathcal{D}}\{E_x\{(h_g^{\mathcal{D}}(x) - \underbrace{\bar{g}(x)}_{0})^2\}\} = E_{\mathcal{D}}\{E_x\{(ax+b)^2\}\}$$
$$= E_{\mathcal{D}}\left\{\frac{a^2}{2}\int_{-1}^{1} x^2\,dx + ab\int_{-1}^{1} x\,dx + b^2\right\} = E_{\mathcal{D}}\left\{\frac{a^2}{3} + b^2\right\}$$
$$= \frac{1}{3}E_{\mathcal{D}}\{(x_1+x_2)^2\} + E_{\mathcal{D}}\{x_1^2x_2^2\}$$
$$= \frac{1}{3}\left(\frac{1}{2}\int_{-1}^{1} x_1^2\,dx_1 + 2\cdot\frac{1}{4}\int_{-1}^{1}\int_{-1}^{1} x_1+x_2\,dx_1\,dx_2 + \frac{1}{2}\int_{-1}^{1} x_2^2\,dx_2\right) + \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1} x_1^2x_2^2\,dx_1\,dx_2$$
$$= \frac{1}{3}\left(\frac{1}{3}+\frac{1}{3}\right) + \frac{1}{9} = \frac{1}{3}$$

2. AML Problem 4.4 (a)-(c), plus additional parts (i)-(iii) below.

Problem 4.4 LAMi This problem is a detailed version of Exercise 4.2. We set up an experimental framework which the reader may use to study various aspects of overfitting. The input space is $\mathcal{X} = [-1, 1]$, with uniform input probability density, $P(x) = \frac{1}{2}$. We consider the two models $\mathcal{H}_2$ and $\mathcal{H}_{10}$. The target function is a polynomial of degree Q_f, which we write as $f(x) = \sum_{q=0}^{Q_f} a_q L_q(x)$, where $L_q(x)$ are the Legendre polynomials. We use the Legendre polynomials because they are a convenient orthogonal basis for the polynomials on $[-1, 1]$ (see Section 4.2 and Problem 4.3 for some basic information on Legendre polynomials). The data set is $\mathcal{D} = (x_1, y_1), \ldots, (x_N, y_N)$, where $y_n = f(x_n) + \sigma\epsilon_n$ and ϵ_n are iid standard Normal random variates.

For a single experiment, with specified values for Q_f, N, σ, generate a random degree-Q_f target function by selecting coefficients a_q independently from a standard Normal, rescaling them so that $\mathbb{E}_{\mathbf{a},x}\left[f^2\right] = 1$. Generate a data set, selecting $x_1, \ldots, x_N$ independently from $P(x)$ and $y_n = f(x_n) + \sigma\epsilon_n$. Let g_2 and g_{10} be the best fit hypotheses to the data from $\mathcal{H}_2$ and $\mathcal{H}_{10}$ respectively, with respective out of-sample errors $E_{\text{out}}(g_2)$ and $E_{\text{out}}(g_{10})$.

(a) Why do we normalize f? *[Hint: how would you interpret σ?]*

(b) How can we obtain g_2, g_{10}? *[Hint: pose the problem as linear regression and use the technology from Chapter 3.]*

(c) How can we compute E_{out} analytically for a given g_{10}?

>> For part (c), assume both $g_{10}(x)$ and $f(x)$ are given as functions of x, and you can express your answer in terms of them; and define $E_{out}(g_{10}) = \mathrm{E}_{x,y}\left\{\left[g_{10}(x) - y(x)\right]^2\right\}$.

(i) In Fig. 4.3(a), set $\sigma^2 = 0.5$, and traverse the horizontal line from N $\approx$ 60 to N $\approx$ 130. Explain why $\mathcal{H}_{10}$ transitions from overfit to good fit (relative to $\mathcal{H}_2$).

(ii) Also in Fig. 4.3(a), set N = 100, and traverse the vertical line from $\sigma^2 = 0$ to $\sigma^2 = 2$. Explain why $\mathcal{H}_{10}$ transitions from good fit to overfit (relative to $\mathcal{H}_2$).

(iii) In Fig. 4.3(b), set N $\approx$ 75, and traverse the vertical line from $Q_f = 0$ to $Q_f = 100$. Explain the behavior.

(a) When the "signal" f is normalized to $\mathbb{E}[f^2] = 1$, the noise level σ^2 is automatically calibrated to the signal level.

(b) To obtain g_2, we transform the original data with a second order transformation.

$$g_2(x) = \tilde{w}^T \underline{x} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}^T \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}^T \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} = \tilde{w}^T \Phi_2(x)$$

Then, we can find the best linear fit $\tilde{w}$ and use $g_2(x) = \tilde{w}\underline{x}$ to get $g_2(x)$.

Also, to obtain g_{10}, we need to transform the original data with a tenth order transformation

$$g_{10}(x) = \tilde{w}^T \underline{x} = \tilde{w}^T \Phi_{10}(x)$$

Then, we can find the best linear fit $\tilde{w}$ and use $g_{10}(x) = \tilde{w}\underline{x}$ to get $g_{10}(x)$

(c)

$$E_{out}(g_{10}) = \mathbb{E}_{x,y}\{[g_{10}(x) - y(x)]^2\} = \mathbb{E}_{x,y}\{[g_{10}(x) - f(x) - \sigma\epsilon]^2\}$$

$$= \mathbb{E}_{x,y}\{(g_{10}(x) - f(x))^2 - 2\sigma\epsilon(g_{10}(x) - f(x)) + \sigma^2\epsilon^2\}$$

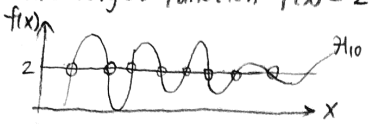
$\because \epsilon_n$ are IID standard normal random RV.

$$= \mathbb{E}_{x,y}\{(g_{10}(x) - f(x))^2\} + \sigma^2$$

(i) When N is small, $\mathcal{H}_{10}$ will try its best to fit every sample in traing data, therefore, cause severe overfitting. However, when N becomes large, since the degree of $\mathcal{H}$ is limited, it's hard to fit the whole samples, it can only find a curve that fits the samples best which won't cause the overfitting problem.

(ii) Since machine cannot tell if the output function contains noise or not, when the stochastic noise is large, $\mathcal{H}$ will try to fit the points that have large stochastic noise, therefore, cause the overfitting problem.

(iii) At first, the color is red which means it suffers from serious overfitting. because the target is so simple, and we use a complexed $\mathcal{H}$ to fit.
e.g. the target function $f(x) = 2$, and the $\mathcal{H}$ is tenth order. ⇒ overfitting



Then, as we move up, the target function becomes more complicated that our $\mathcal{H}_{10}$ fits the samples well. We can find that if we keep moving up, the deterministic noise increases as Q_f increases. which leads to bigger E_{out} and overfitting.

Exercise 4.5 [Tikhonov regularizer]

A more general soft constraint is the *Tikhonov* regularization constraint

$$\mathbf{w}^{\mathrm{T}}\Gamma^{\mathrm{T}}\Gamma\mathbf{w} \leq C$$

which can capture relationships among the w_i (the matrix Γ is the Tikhonov regularizer).

(a) What should Γ be to obtain the constraint $\sum_{q=0}^{Q} w_q^2 \leq C$?

(b) What should Γ be to obtain the constraint $(\sum_{q=0}^{Q} w_q)^2 \leq C$?

(a) $\underline{w}^T \underline{\underline{\Gamma}}^T \underline{\underline{\Gamma}} \underline{w} = \Sigma_{q=0}^{Q} w_q^2 \Rightarrow \underline{\underline{\Gamma}}$ is an identity matrix $I_{(Q+1)\times(Q+1)}$

(b) $\underline{w}^T \underline{\underline{\Gamma}}^T \underline{\underline{\Gamma}} \underline{w} = (\Sigma_{q=0}^{Q} w_q)^2 \Rightarrow \underline{w}^T \underline{\underline{\Gamma}}^T = \Sigma_{q=0}^{Q} w_q$

$\Rightarrow \underline{\underline{\Gamma}}^T = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{(Q+1)\times 1} \Rightarrow \underline{\underline{\Gamma}} = [1\ 1\ \cdots\ 1]_{1\times(Q+1)}$