Problem 1.12

(a) In sample mean:

We have the equation

$$E_{in}(h) = \sum_{n=1}^{N} (h - y_n)^2 = f(h)$$

Taking derivative and set the derivative to zero:

$$\frac{df(h)}{dh} = \sum_{n=1}^{N} 2(h - y_n) = 0$$

We then have

$$2\sum_{n=1}^{N} h - 2\sum_{n=1}^{N} y_n = 0 \rightarrow Nh = \sum_{n=1}^{N} y_n$$

$$h = \frac{1}{N}\sum_{n=1}^{N} y_n = h_{mean}$$

That is, the estimate will be the in-sample mean.

(b) In sample median:

We have the equation

$$E_{in}(h) = \sum_{n=1}^{N} |h - y_n|$$

Assume that y_n is sorted in ascending order

Suppose that for some integer $k \in [1, N]$ such that $y_i < h\, for\ i < k$ and $y_j \geq h\, for\ j \geq k$

$$E_{in}(h) = \sum_{i=1}^{k-1} (h - y_i) + \sum_{j=k}^{N} (y_j - h)$$

$$E_{in}(h) = \sum_{i=1}^{k-1} (h - y_i) - \sum_{j=k}^{N} (h - y_j) = f(h)$$

Taking derivative and set the derivative to zero:

$$\frac{df(h)}{dh} = \sum_{i=1}^{k-1} 1 - \sum_{j=k}^{N} 1 = 0 \rightarrow k = \frac{N}{2}$$

Because k is equal to half N, then $h = h_{med}$.

That is, to minimize the absolute deviation, the estimate will the median because half the data points are at most h_{med} and half the data points are at least h_{med}.

(c) Outliners:

As $\varepsilon \to \infty$, $h_{mean} \to \infty$ because we have $h_{mean} = \frac{1}{N}\sum_{n=1}^{N} y_n$ and hence h_{mean} is affected by all the data points including the outliners.

As $\varepsilon \to \infty$, h_{med} does not change because median is only influenced by the middle two data points. Perturbed data point y_n will have no influence on the h_{med}.

Problem 2.3

(a) The dichotomy generated by negative ray is not same as the dichotomy generated by positive ray except two cases, which are when all data points are positive, and when all data points are negative.

Therefore the most number of dichotomies is $m_H(N) = 2 \times (N + 1) - 2(double\ counting) = 2N$

And setting the equation $2N \leq 2^N$ to obtain the maximum $N = 2$, hence the VC-Dimension $d_{vc} = 2$

(b) The dichotomy generated by negative interval is not the same as the dichotomy generated by positive interval except two scenarios, which are when the first n data points are negative, and when the last n data points are negative.

Because $n \leq N$, The number of double counting is therefore $2N$

Therefore the most number of dichotomies is

$$m_H(N) = 2 \times \left(\frac{(N+1)N}{2} + 1\right) - (2N) = N^2 - N + 2$$

And setting the equation $N^2 - N + 2 \leq 2^N$ to get minimum $N = 3$, hence the VC-Dimension $d_{vc} = 3$

(c) For the concentric spheres, it is very similar to the positive intervals since we compute the norm and match it with the interval $[a, b]$.

Therefore the maximum number of dichotomies is $m_H(N) = \frac{(N+1)N}{2} + 1$ and $d_{vc} = 2$

Problem 2.8

Possible growth functions:

$1 + N$ example: positive rays will generate $N + 1$ dichotomy for N data points. Therefore the growth function is $1 + N$.

$1 + N + \frac{N(N-1)}{2}$ example: positive interval. We need to choose two places for the interval start and end from $N + 1$ possible locations. Therefore the growth function is $C_2^{N+1} = \binom{N+1}{2} = 1 + N + \frac{N(N-1)}{2}$

2^N example: convex set. Any dichotomy on the N points can be realized using a convex hypothesis; therefore the growth function is 2^N.

$2^{\lfloor\sqrt{N}\rfloor}$ is a possible growth function because for all $N > 0$, $2^{\lfloor\sqrt{N}\rfloor} < 2^N$

$2^{\lfloor N/2\rfloor}$ is a possible growth function similarly because for all $N > 0$, $2^{\lfloor\sqrt{N}\rfloor} < 2^N$

Impossible growth functions:

$1 + N + N(N-1)(N-2)/6$:

When $N = 2$, $1 + N + N(N-1)(N-2)/6 = 3 < 2^N$

When $N = 1$, $1 + N + N(N-1)(N-2)/6 = 2 = 2^N$

This contradicts the definition of d_{vc} that all $N < d_{vc} + 1$ will have $m_H(N) < 2^N$

Problem 2.10

Let the most number of dichotomies can be generated on N data points be k.we have $m_H(N) = k$.

Consider $2N$ data points is partitioned into two set of data points of equal size, then for each set of data points the most number of dichotomies can be generated is k.

If we now combine the two sets back to a single set of $2N$ data points, then the must number of dichotomies possible will be the cross product of the most dichotomies of two sets. That is

$$m_H(2N) \leq k \times k = m_H(N)^2$$

Problem 2.13

(a) By definition: $m_H(N) = \max_{x_1,\dots,x_n \in X} |H(x_1, \dots, x_n)|$

Since $H = \{h_1, h_2, \dots, h_M\}$, we have $|H| = M$

Because it is possible that different hypothesis in H generate same dichotomy on a given data set X. We have the inequality:

$$m_H(N) \leq M$$

Because the M is finite, and by definition the VC dimension of a hypothesis is the largest value of N for which $m_H(N) = 2^N$

That is:

$d_{vc} = N$ such that $m_H(N) = 2^N$

Because $m_H(N) \leq M$, we have $m_H(N) = 2^N \leq M \to d_{vc} \leq \log_2 M$

(b)

$$0 \leq d_{vc}\left(\bigcap_{k=1}^{K} H_k\right) \leq \min_{k\in[1,K]}\{d_{vc}\,(H_k)\}$$

For the lower bound, it is possible that the intersection of the hypothesis set is an empty set. Therefore the smallest possible VC dimension is zero.

For the upper bound, in part (a) we proved that for finite M, $d_{vc} \leq \log_2 M$. Therefore hypothesis from higher d_{vc} hypothesis set will be filtered out because they are not contained in lower d_{vc} hypothesis set because $d_{vclow} < d_{high}$ implies $M_{low} < M_{high}$.

The best case happens when the hypothesis in the minimum d_{vc} hypothesis set is contained in all other hypothesis sets. Otherwise, the VC dimension of the intersection will only become lower.

(c)

$$\max_{k\in[1,K]}\{d_{vc}\,(H_k)\} \leq \; d_{vc}\left(\bigcup_{k=1}^{K} H_k\right) \leq \sum_{k=1}^{K} d_{vc}(H_k) + K - 1$$

Worst cases happens when hypothesis from lower d_{vc} hypothesis sets are all contained in the higher d_{vc} hypothesis set. Therefore the union will not increase the VC dimension.

Let $H_{1,2} = H_1 \cup H_2$

The number of dichotomies of N particular points can be generated using $H_{1,2}$ is at most the number of dichotomies using H_1 + the number of dichotomies using H_2. We therefore have:

$$m_{H_{1,2}}(N) \le m_{H_1}(N) + m_{H_2}(N) \rightarrow m_{H_{1,2}}(N) \le \sum_{i=0}^{d_{vc}(H_1)} \binom{N}{i} + \sum_{i=0}^{d_{vc}(H_2)} \binom{N}{i}$$

Because $\binom{N}{i} = \binom{N}{N-i}$, we have:

$$m_{H_{1,2}}(N) \le \sum_{i=0}^{d_{vc}(H_1)} \binom{N}{i} + \sum_{i=0}^{d_{vc}(H_2)} \binom{N}{N-i} \le \sum_{i=0}^{d_{vc}(H_1)} \binom{N}{i} + \sum_{i=N-d_{vc}(H_2)}^{N} \binom{N}{N-i}$$

Let $N - d_{vc}(H_2) > d_{vc}(H_1) + 1$, which is $N \ge d_{vc}(H_1) + d_{vc}(H_2) + 2$

$$m_{H_{1,2}}(N) \le \sum_{i=0}^{N} \binom{N}{i} - \binom{N}{d_{vc}(H_1) + 1} = 2^N - \binom{N}{d_{vc}(H_1) + 1} \le 2^N$$

Therefore:

$$d_{vc}(H_{1,2}) \le d_{vc}(H_1) + d_{vc}(H_2) + 1$$

Apply this inequality to the union of multiple hypothesis set we obtain that

$$d_{vc}\left(\bigcup_{k=1}^{K} H_k\right) \le \sum_{k=1}^{K} d_{vc}(H_k) + K - 1$$

Problem 2.22

Given in the textbook:

$$E_D\left[E_{out}\left(g^{(D)}\right)\right] = E_D\left[E_x\left[\left(g^{(D)}(x) - y(x)\right)^2\right]\right]$$

$$= E_x\left[E_D\left[g^{(D)}(x)^2\right] - 2E_D\left[g^{(D)}(x)\right]y(x) + y(x)^2\right]$$

Denote the average function $E_D\left[g^{(D)}(x)\right] = \bar{g}(x)$, the text book decompose the equation into:

$$E_D\left[E_{out}\left(g^{(D)}\right)\right] = E_x\left[E_D\left[\left(g^{(D)}(x) - \bar{g}(x)\right)^2\right] + \left(\bar{g}(x) - y(x)\right)^2\right]$$

Because $var(x) = E_D\left[\left(g^{(D)}(x) - \bar{g}(x)\right)^2\right]$

$$E_D\left[E_{out}\left(g^{(D)}\right)\right] = E_x[var(x)] + E_x\left[\left(\bar{g}(x) - y(x)\right)^2\right]$$

$$= var + E_x\left[\left(\bar{g}(x) - y(x)\right)^2\right]$$

Because $y(x) = f(x) + \varepsilon$

$$E_x\left[\left(\bar{g}(x) - y(x)\right)^2\right] = E_x[(\bar{g}(x) - f(x) - \varepsilon)^2]$$

$$= E_x[\bar{g}(x)^2 - 2\bar{g}(x)f(x) - 2\varepsilon\bar{g}(x) + f(x)^2 - 2\varepsilon f(x) + \varepsilon^2]$$

$$= E_x\left[\left(\bar{g}(x) - f(x)\right)^2\right] + E_x\left[-2\varepsilon\left(f(x) - \bar{g}(x)\right) + \varepsilon^2\right]$$

Because $\bar{g}(x)$ is the average function and ε is a zero-mean random noise

$$E_x\left[\left(\bar{g}(x) - y(x)\right)^2\right] = bias + -2\varepsilon E_x\left[\left(f(x) - \bar{g}(x)\right)\right] + E_x[\varepsilon^2]$$

$$= bias + 0 + \sigma^2$$

That is:

$$E_D\left[E_{out}\left(g^{(D)}\right)\right] = var + bias + \sigma^2$$

Problem 2.24

(a) the average function $\bar{g}(x)$ will have the average slope and average constant term

$$\bar{g}(x) = \frac{1}{N}\sum_{n=1}^{K} g_n(x) = \left(\frac{1}{N}\sum_{n=1}^{K} a_n\right)x + \frac{1}{N}\sum_{n=1}^{K} b_n$$

Because

$$\bar{a} = \frac{1}{N}\sum_{n=1}^{K} a_n = \frac{1}{N}\sum_{n=1}^{K}(x_{n1} + x_{n2})$$

$$\bar{b} = \frac{1}{N}\sum_{n=1}^{K} b_n = \frac{1}{N}\sum_{n=1}^{K}(x_{n1}x_{n2})$$

Then we have

$$\bar{g}(x) = \bar{a}x + \bar{b}$$

(b) randomly generate 1000 data set and calculate the function $g(x)$ for each dataset.

```
for times = 1:1000
    x1 = rand*2-1;
    y1 = x1.^2;

    x2 = rand*2-1;
    y2 = x2.^2;

    data = [data; [x1,y1],[x2,y2]];

    coeff = polyfit([x1,x2],[y1,y2],1);

    slope = [slope; coeff(1)];
    constant = [constant; coeff(2)];
end
```

Compute average function $\bar{g}(x)$, the variance, the bias, and the out of sample error.

```
       avgSlope = mean(slope);
       avgConstant = mean(constant);

       for times = 1:1000
           bias = bias + (avgSlope*data(times,1)+avgConstant-data(times,1)).^2;
           bias = bias + (avgSlope*data(times,1)+avgConstant-data(times,3)).^2;

           variance = variance + mean((avgSlope*data(times,1)+avgConstant-
data(times,2)).^2);
           variance = variance + mean((avgSlope*data(times,1)+avgConstant-
data(times,4)).^2);
       end

       bias = bias/1000;
       variance = variance / 1000;
```

(c)

$$\bar{g}(x) = 0.021x + 0.004$$

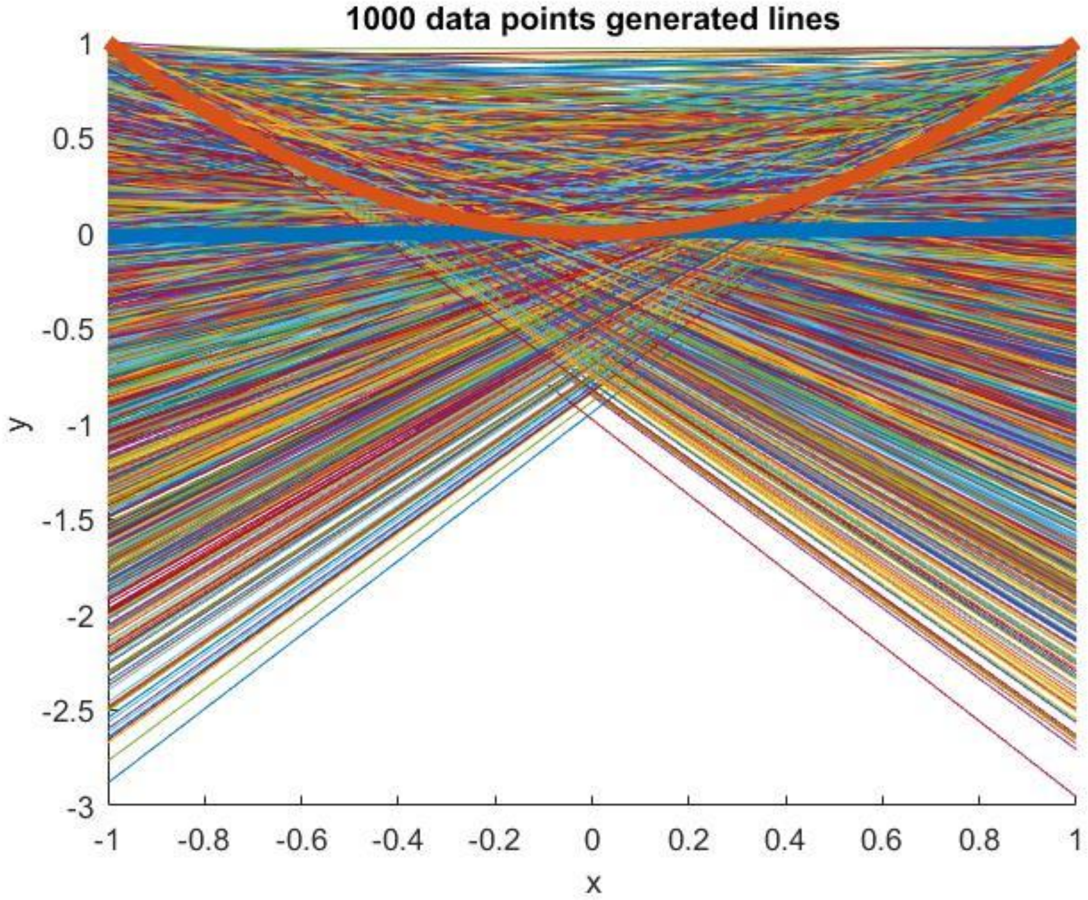
$$bias = 0.201$$

$$variance = 0.322$$

$$E_{out} = 0.527$$

As proved in problem 2.22, $E_{out} = variance + bias$ when there is no noise in the data. And the experiment run results matches the equation: $0.201 + 0.322 = 0.523 \approx 0.527$

In the graph, the thick orange line is $y = x^2$ and the thick blue line is $\bar{g}(x)$.



(d)

$$E_{out} = \int_{-1}^{1} \frac{1}{2}[g(x) - f(x)]^2 dx = \int_{-1}^{1} \frac{1}{2}[ax + b - x^2]^2 dx$$

$$E_{out} = \int_{-1}^{1} \frac{1}{2}[x^4 - 2ax^3 + (a^2 - 2b)x^2 + 2abx + b^2]dx = \frac{1}{5} + \frac{a^2 - 2b}{3} + b^2$$
$$= \frac{1}{5} + \frac{1}{3}E[(x_1 + x_2)^2] - \frac{2}{3}E[x_1x_2] + E[(x_1x_2)^2]$$

Because we have

$$E[(x_1 + x_2)^2] = E[x_1^2] + 2E[x_1]E[x_2] + E[x_2^2] = \frac{1}{3} + 0 + \frac{1}{3} = \frac{2}{3}$$

$$E[x_1x_2] = E[x_1]E[x_2] = 0$$

$$var(x_1) = var(x_2) = E[x_1^2] - E[x_1]^2 = E[x_2^2] - E[x_2]^2 = \frac{1}{3} - 0 = \frac{1}{3}$$

$$var = \frac{1}{2}\big(var(x_1) + var(x_2)\big) = \frac{1}{3}$$

Then we have

$$E_{out} = \frac{1}{5} + \frac{1}{3}\frac{2}{3} - 0 + \frac{1}{3}\frac{1}{3} = \frac{8}{15} = 0.53\dot{3}$$

Because $\bar{a} = \bar{b} = 0$

$$bias = \int_{-1}^{1} \frac{1}{2}[\bar{g}(x) - f(x)]^2 dx = \frac{1}{5} + \frac{\bar{a}^2 - 2\bar{b}}{3} + \bar{b}^2 = \frac{1}{5}$$

And we have

$$var + bias = \frac{1}{3} + \frac{1}{5} = \frac{8}{15} = E_{out}$$