Exercise 2.8

a)

Since we have $\overline{g}(x) \approx \frac{1}{K}\sum_{k=1}^{K} g_k(x)$,we could conclude that $\overline{g}(x)$ is a linear combination of K final hypothesis from H based on K data sets.

Since H is closed under linear combination, $\overline{g} \in H$.

b)

$H = \{g_1(x), g_2(x)\}$

$g_1(x) = 2,\ g_2(x) = 1$

$\overline{g}(x) = \frac{2+1}{2} = 1.5$, this does not exist in the hypothesis set.

c)

Not really.

The only way the $\overline{g}$ can be binary function if that all of $g_k(x)$ produce 1 or all of $g_k(x)$ produce 0. In both cases, the $\overline{g}(x)$ can be 1 or 0.

In other cases, the $\overline{g}(x)$ will be equal to (number of 1s / K), which is not a binary result. Hence, we could conclude that we could not expect $\overline{g}$ to be a binary function.

Problem 2.14

a)

We have $H = \bigcup_{k=1}^{K} H_k$.

We consider the worst case that these hypotheses do not have intersection between each other.

We have $d_{vc}(H) = \sum_{k=1}^{K} d_{vc} = Kd_{vc}$. $d_{vc} + 1$ is the breakpoint for these hypotheses.

We could conclude that $d_{vc}(H) < K(d_{vc} + 1)$.

If for all the hypotheses in H, $M_H(N) = 2^N$, the growth function of H is also $M_H(N) = 2^N$, so we have the $d_{vc}(H) = \infty$. $\infty < (\infty + 1)$. $d_{vc}(H) < K(d_{vc} + 1)$ is also true.

b)

We have $M_H(N) \leq N^{d_{vc}} + 1$

We can deduce this $M_H(N) \leq N^{d_{vc}} + 1 \leq 2N^{d_{vc}}$ by using the above step.

Also, from the hw4 problem 2.3, we can also conclude that

$M_{H1UH2}(N) \leq M_{H1}(N) + M_{H2}(N)$

In this case, we can have $M_{H1UH2...UHk}(N) \leq \sum_{i=1}^{K} M_{Hi}(N)$

According to the condition given by the problem, we can have

$M_H(l) = M_{H1UH2...UHk}(l) \leq \sum_{i=1}^{K} M_{Hi}(N) \leq \sum_{i=1}^{K} 2l^{d_{vc}} = 2Kl^{d_{vc}}$

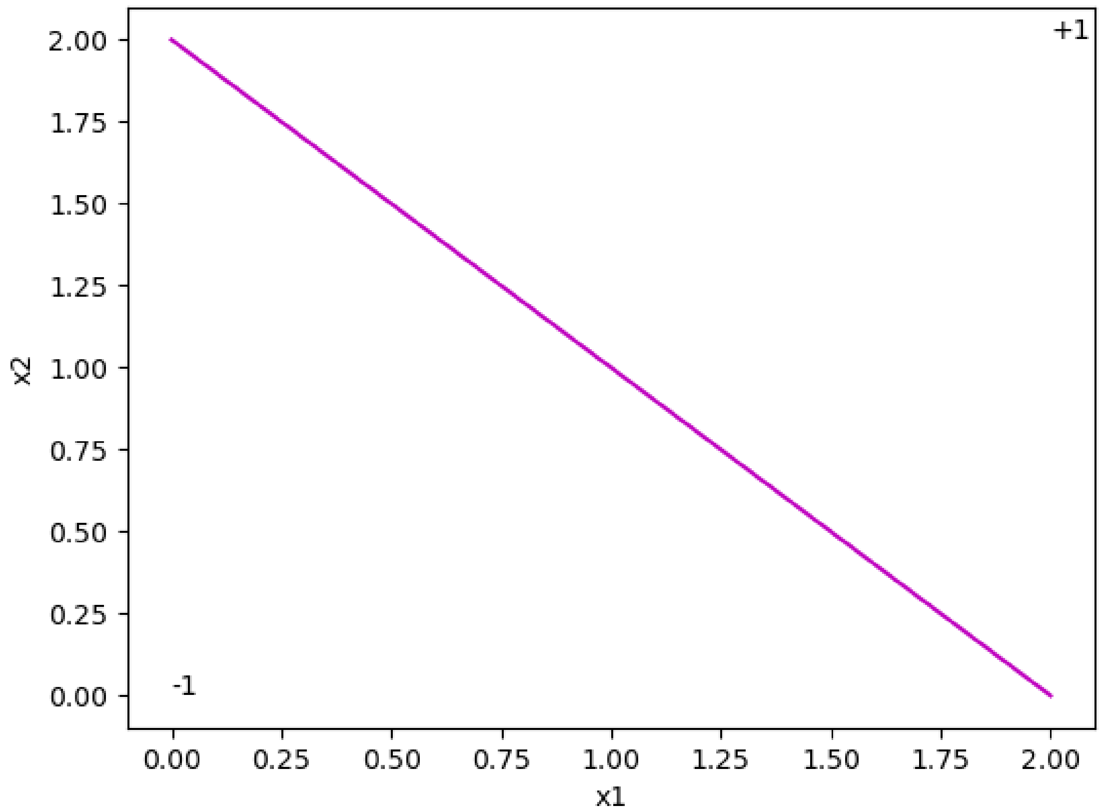
We have $2^l < 2Kl^{dvc}$ from the problem

$M_H(l) <= 2Kl^{d_{vc}} < 2^l$

Then we will have $d_{vc}(H) < l$

Problem 2.15

a)



b)

In 1D, the problem is similar to the positive ray problem, since the larger valued points need to have larger h(x). The $M_H(N) = N+1$ and $d_{vc}(H) = 1$

Using the hint, in 2D, we construct N points where we pick an arbitrary point and create the following points by increasing the first component and decreasing the second component of the previous point.

We could observe that there is no point that could be larger than the other one. In this case, no matter how the target y changes, H is able to classify them; because the points in the data set are not constrained by the rule. We are able to construct a similar set of points in any dimension.

Hence, we can conclude that H can shatter all the points. $M_H(N) = 2^N$ and $d_{vc}(H) = \infty$.

Problem 2.24

a)

In order to minimize the E_{in}, we choose the hypothesis line which passes through the $(x1,x1^2)$ and $(x2, x2^2)$.

$g(x) = ax + b$

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

$b = - x_1 x_2$

$\bar{g}(x) = E_D[g^{(D)}(x)]$

$\bar{g}(x) = E_D[(x_1 + x_2)x - x_1 x_2]$

$\bar{g}(x) = E_D[x_1]x + E_D[x_2]x - E_D[x_1]E_D[x_2]$ the independence between x_1 and x_2.

Since we have a uniform distributed input space with interval [-1,+1], $E_D[x_1] = E_D[x_2] = 0$.

Hence, we can conclude that $\bar{g}(x) = 0$.

b)

Generate K numbers (ex: 5000 or more) of uniformed distributed data sets.

$\bar{g}(x)$: Iterate through these data sets and compute the average a $(x_2 + x_1)$ and average b $(- x_1 x_2)$. We can finally get $\bar{g}(x)$.

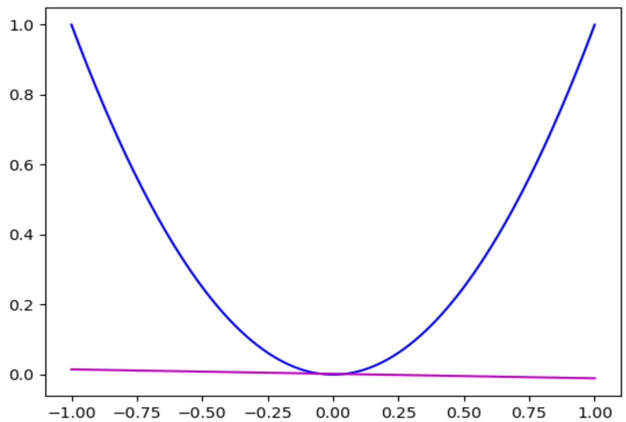
bias: we have $bias(x) = (\bar{g}(x) - f(x))^2$, $bias = E_x[bias(x)]$, we can compute it by getting the average of the bias generated on each data set.

var: we have $var(x) = E_D[(g^{(D)}(x) - \bar{g}(x))^2] = E_D[g^{(D)}(x)^2] - \bar{g}(x)^2$, $var = E_x[var(x)]$, we need to first compute the $E_D[g^{(D)}(x)^2]$, which is the average of $g^{(D)}(x)^2$ based on the K datasets we have. After getting $E_D[g^{(D)}(x)^2]$, we need to compute $E_x[E_D[g^{(D)}(x)^2]]$, which is the average of $E_D[g^{(D)}(x)^2]$ based on the input space. To achieve this, we need to generate K number of uniform distributed number from -1 to +1 and we plug it in to the $E_D[g^{(D)}(x)^2]$ and divided by K.

$E_D[E_{out}(g^{(D)})]$**:** we have $E_x[E_D[g^{(D)}(x)^2] - 2E_D[g^{(D)}(x)]f(x) + f(x)^2]$ from the book. We can add the terms we get from previous var and $\bar{g}(x)$. Then, we compute the average of the addition of the term by performing the final step of computing var.

c)

The blue line is f(x) and the purple line is $\bar{g}(x)$




```
bias: 0.20
Var: 0.34
E out: 0.54
E out: 0.54 by adding bias and var
```

After computing, we got the bias about 0.20 and variance about 0.34; E_{out} is 0.54, which is similar to the addition of var and bias.

d)

$E_{out}(g^{(D)})$:

$E_{out}(g^{(D)})$ $=E_x[(g^{(D)}(x) - f(x))^2]$

$E_{out}(g^{(D)})$ $=E_x[(ax + b - x^2)^2]$

$E_{out}(g^{(D)})$ $=E_x[x^4] - 2aE_x[x^3] + (a^2-2b)E_x[x^2] + 2abE_x[x] + b^2$

$E_{out}(g^{(D)}) = \frac{1}{2}\int_{-1}^{1} x^4 dx - 2a \times \frac{1}{2}\int_{-1}^{1} x^3 dx + (a^2 - 2b) \times \frac{1}{2}\int_{-1}^{1} x^2 dx + 2ab \times \frac{1}{2}\int_{-1}^{1} x dx$ $+ b^2$

$E_{out}(g^{(D)}) = \frac{1}{5} + \frac{1}{3}(a^2 - 2b) + b^2$

$E_D[E_{out}(g^{(D)})]$:

$E_D[E_{out}(g^{(D)})] = E_D[\frac{1}{5} - \frac{1}{3}(a^2 - 2b) + b^2]$

$= \frac{1}{5} - \frac{1}{3}E_D[a^2 - 2b] + E_D[b^2]$

$= \frac{1}{5} - \frac{1}{3} \times \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1}(x_1 + x_2)^2 - 2x_1x_2dx_1dx_2 + \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1} x_1^{\ 2}x_2^{\ 2}dx_1dx_2$

$= \frac{1}{5} + \frac{1}{3} \times \frac{1}{4} \times \frac{8}{3} + \frac{1}{4} \times \frac{4}{9} = \frac{8}{15} \approx 0.53$, which is close to the 0.54 from the program

Bias:

$bias = E_x[(\overline{g}(x) - f(x))^2]$

$= E_x[(x^2)^2]$

$= E_x[x^4]$

$bias = \frac{1}{2}\int_{-1}^{1} x^4dx = \frac{1}{5} = 0.2$

Var:

$var(x) = E_D[(g^{(D)}(x) - \overline{g}(x))^2]$

$= E_D[(ax + b)^2] = E_D[a^2x^2 + 2axb + b^2]$

$= E_D[a^2]x^2 + 2E_D[ab]x + E_D[b^2]$

$= \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1} x_1^{\ 2} + 2x_1x_2 + x_2^{\ 2}dx_1dx_2 \times x^2 - \frac{2}{4}\int_{-1}^{1}\int_{-1}^{1} x_1^{\ 2}x_2 + x_1x_2^{\ 2}dx_1dx_2 \times x + \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1} x_1^{\ 2}x_2^{\ 2}dx_1d$

$= \frac{2}{3}x^2 - 0 + \frac{1}{9} = \frac{2}{3}x^2 + \frac{1}{9}$

$var = E_x[E_D[(g^{(D)}(x) - \overline{g}(x))^2]]$

$= E_x[\frac{2}{3}x^2 + \frac{1}{9}]$

$= \frac{2}{3} \times \frac{1}{2}\int_{-1}^{1} x^2dx + \frac{1}{9}$

$= \frac{2}{3} \times \frac{1}{3} + \frac{1}{9} = \frac{1}{3} \approx 0.33$, which is close to the 0.34 from the program.