1. Exercise 2.8

   (a) We know that $\overline{g}$ is the average function of many different hypothesis $g_1,\ g_2, ...,\ g_n$ of different data sets. H represents the hypotheses set, where each hypothesis in H is dependent on their respective data set. If the hypothesis in H are in linear combination, then the average of the hypotheses in H should also be a linear combination, proving that $\overline{g} \in H$ ■

   (b) We can imagine a model with two hypotheses, one that will label all datapoints as $+1$ and another as -1, then the average of those two hypotheses will be 0, which is not in the hypothesis set, therefore $\overline{g} \notin H$ ■

   (c) (b) is a binary classification and $\overline{g}$ is not a binary function since $0 \notin \{+1, -1\}$. Often, with more hypotheses, the average will be a number between $\{-1, 1\}$, it will be unlikely that they are $+1$ or -1

2. Problem 2.14

   (a) Given that the d_{vc} is finite, we know that the hypothesis can shatter any data set of size d_{vc}. We can claim that $K(d_{vc}) \geq d_{vc}(H)$, since $K(d_{vc})$ assumes that every K hypothesis can shatter the maximum number of points, d_{vc}. Then $K(d_{vc}+1) > K(d_{vc}) \geq d_{vc}(H)$, by transitivity, $K(d_{vc}+1) > d_{vc}(H)$ ■

   (b) Given that the hypothesis has a finite VC dimension, then we have a breakpoint such that $2^\ell > \ell^{d_{vc}} + 1$ If we have K hypotheses, then $m_H(\ell) \leq K(\ell^{d_{vc}} + 1)$. By inspection, $2K\ell^{d_{vc}} > K(\ell^{d_{vc}} + 1)$. So here we know that $2^\ell > 2K\ell d^{vc} > K(\ell^{d_{vc}} + 1)$, by transitivity, $2^\ell > K(\ell^{d_{vc}} + 1)$, which means $2^\ell > m_H(\ell)$, if this is true, then it means that there are no hypothesis can shatter 2^ℓ points, which means that $d_{vc} \leq \ell$ ■

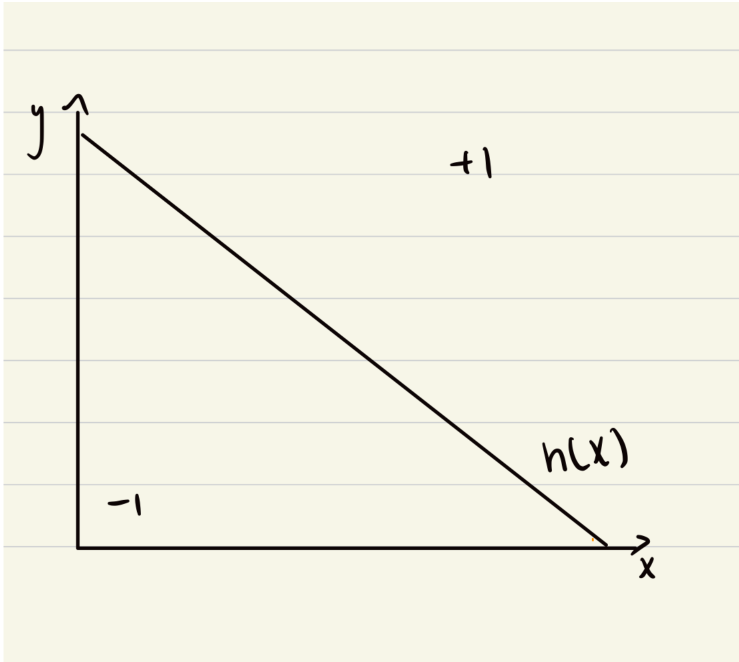
   (c) From (a), we see $K(d_{vc}+1) > d_{vc}(H)$, so if we assume $K(d_{vc}+1)$ is the min, we know that the inequality statement will be true. To prove the second part of the min, we let $\ell = 7(d_{vc}+K)\log_2(d_{vc}K) = \log_2(d_{vc}K)^{7(d_{vc}+K)}$, so:

   $$\begin{aligned} 2^{\log_2(d_{vc}K)^{7(d_{vc}+K)}} &= d_{vc}K^{7(d_{vc}+K)} \\ d_{vc}K^{7(d_{vc}+K)} &> 2K(7(d_{vc}+K)\log_2(d_{vc}K))^{d_{vc}} \end{aligned}$$

   We know that we have a finite VC dimension, but our number of hypothesis k has no bound and can grow infinitely, therefore we prove the inequality is true.

3. Problem 2.15

   (a) Consider x_1 is a point with coordinates i_1, j_1, then $h(i_1, j_1) = sign(i_1 + j_1)$. For the purpose of this example, we will consider $x_2 = (i_2, j_2)$, where $h(i_2, j_2) = sign(i_2, j_2)$. If $i_1 > i_2$ and $j_1 > j_2$, then $sign(i_1 + j_1) > sign(i_2 + j_2)$. We can consider the following graph:



(b) If we begin with a point and increase the first component, let that be i_1 and decrease the second component, let that be j_1, then we will continuously generate the value $sign(i_n, j_n)$ for $\sum_{n=1}^{N}$. It doesn't matter the number of points that are generated, there will always be a way to represents all the dichotomies, therefore, the maximum number of dichotomies is $m_H(N) = 2^N$, by definition, $d_{VC} = \infty$

4. Problem 2.24

(a) We don't know the target function, but out hypothesis set has functions in the form of $ax+b$, so $x_1^2 = ax_1+b$ and $x_2^2 = ax_2+b$, with algebra manipulation, we get to $a = x_1+x_2$ and $b = -x_2x_1$. By definition, we know $\overline{g}(x) = E_D[g^{(D)}(x)] = E_D[(x_1+x_2)x - x_2x_1]$

$$\begin{aligned}\overline{g}(x) &= x \cdot E[(x_1 + x_2 - x_2x_1)] \\ &= x \cdot E[x_1] + E[x_2] - E[x_2x_1]\end{aligned}$$

By definition, $E[x] = \int_{-1}^{+1} x dx = 0$, given that we have two datapoints, the average $\overline{g}(x) = \frac{1}{2}(x \times 0 + 0 - 0) = 0$

(b) Generate a test dataset of size $2N$, then we run N trials, picking two points at a time:

- for $\overline{g}(x)$, generate N trials, each trial contains 2 points, compute $g^D(x) = (x_1+x_2)x - x_2x_1$, then we can compute the average by $\overline{g}(x) = \frac{1}{N}\sum_{i=0}^{N} g_i(x)$
- we can compute out of sample error by (2.17), we generate a new dataset D of length

N, we cannot use the same data we used to find $\overline{g}(x)$

$$\begin{aligned} E_{out}(g^D) &= E_x[(g^D(x) - f(x))^2] \\ &= \frac{1}{N}\sum_{i=1}^{N}(g_i(x_i) - f(x_i))^2 \end{aligned}$$

Consider only the interval from 0 to -1

$$E_{out}(g^D) = \frac{1}{2}\int_{-1}^{1}(\frac{1}{N}\sum_{i=1}^{N}(g_i(x_i) - f(x_i))^2)dx$$

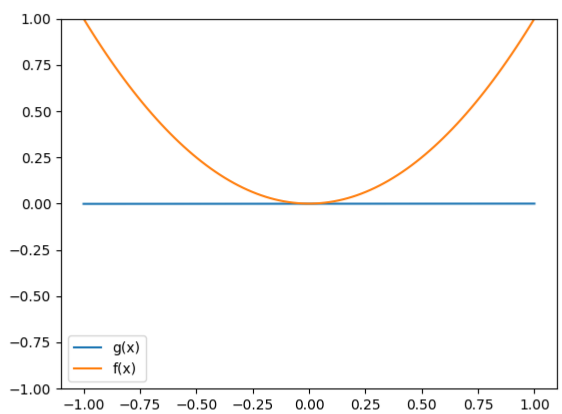
or, if we have bias and var, simply

$$E_D[E_{out}(g^D)] = bias + var$$

- we can compute bias by $bias(x) = \frac{1}{2}\int_{-1}^{1}(\overline{g}(x_i) - f(x_i))^2 dx$
- we can compute var by $var(x) = E_D[(g^D(x) - \overline{g}(x))^2]$

$$E_D[(g^D(x) - \overline{g}(x))^2] = \frac{1}{2}\int_{-1}^{1}(\frac{1}{N}\sum_{i=1}^{N}(g_i(x_i) - \overline{g}(x_i))^2)dx$$

(c) for $N = 1000$, $\overline{g}(x) = -0.0018x + 0.0042$, $E_{out} = 0.5244$, $bias = 0.1891$, $var = 0.3486$, $bias + var = 0.5377$, we had an estimation error $E_{out} \approx bias + var$ by 2.54%



(d) bias: $\frac{1}{2}\int_{-1}^{1}(\overline{g}(x) - f(x))^2 dx$, $\overline{g}(x) = 0$, so $\frac{1}{2}\int_{-1}^{1} x^4 dx = \frac{1}{2} \times \frac{2}{5} = \frac{1}{5}$
var: $E[(g^D(x) - \overline{g}(x))^2]$, since $\overline{g}(x) = 0$, then $var = E[g^D(x)^2]$...???
assume var calculated, then $E_{out} = \frac{1}{5} + var$