Problem Solutions

Chapter 2

Pierre Paquay

Problem 2.1

Let us begin by extracting the value of N from the $\epsilon(M, N, \delta)$ expression. We have that

$$\sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}} \leq \epsilon \Leftrightarrow N \geq \frac{1}{2\epsilon^2}\ln\frac{2M}{\delta}.$$

(a) So for $M = 1$ and $\delta = 0.03$, to have $\epsilon \leq 0.05$ we need

$$N \geq \frac{1}{2 \cdot 0.05^2}\ln\frac{2}{0.03} = 839.9410156.$$

(b) For $M = 100$ and $\delta = 0.03$, to have $\epsilon \leq 0.05$ we need

$$N \geq \frac{1}{2 \cdot 0.05^2}\ln\frac{2 \cdot 100}{0.03} = 1760.9750528.$$

(c) And for $M = 10000$ and $\delta = 0.03$, to have $\epsilon \leq 0.05$ we need

$$N \geq \frac{1}{2 \cdot 0.05^2}\ln\frac{2 \cdot 10000}{0.03} = 2682.00909.$$

Problem 2.2

For $N = 4$, if we consider four non aligned points, this $\mathcal{H}$ shatters these points (you only have to efectively enumerate them to see that all dichotomies can be generated), so in this case we have $m_{\mathcal{H}}(4) = 2^4$.

However, for $N = 5$, no matter how you place your five points, some point will be inside a rectangle defined by others. In this case, we are not able to generate all dichotomies and consequently $m_{\mathcal{H}}(5) < 2^5$.

From these two observations, we may conclude that, for poisitive rectangles, we have $d_{VC} = 4$, thus

$$m_{\mathcal{H}}(N) \leq N^4 + 1.$$

Problem 2.3

(a) We already know that the growth function for positive rays is equal to $N + 1$. If we enumerate the dichotomies added by negative rays, we get $N - 1$ new dichotomies (you get the opposite of the ones from positive rays and you have to substract the two dichotomies where all points are $+1$ and where all points are -1). So in total, we get that

$$m_{\mathcal{H}}(N) = 2N.$$

As the largest value of N for which $m_{\mathcal{H}}(N) = 2^N$ is 2 ($m_{\mathcal{H}}(3) = 6$), we have that $d_{VC} = 2$.

(b) Here, we already know that the growth function for positive intervals is equal to $N^2/2 + N/2 + 1$. If we add the new dichotomies generated by negative intervals, we get $N - 2$ new ones (for example for $N = 3$, we only add the $(+1, -1, +1)$ dichotomy, and for $N = 4$, we add the $(+1, -1, +1, +1)$ and $(+1, +1, -1, +1)$

dichotomies). Of course, this only holds if $N > 1$, in the case where $N = 1$ we already generate the two dichotomies with the positive intervals alone. In conclusion, we may write that

$$m_{\mathcal{H}}(N) = \frac{N^2}{2} + \frac{3N}{2} - 1 \text{ if } N > 1 \text{ and } 2 \text{ if } N = 1.$$

As the largest value of N for which $m_{\mathcal{H}}(N) = 2^N$ is 3 ($m_{\mathcal{H}}(4) = 13$), we have that $d_{VC} = 3$.

(c) To determine the growth function fo concentric circles, we have to map the problem from $\mathbb{R}^d$ to $[0, +\infty[$. To do this, we use the map ϕ defined as

$$\phi : (x_1, \cdots, x_d) \mapsto r = \sqrt{x_1^2 + \cdots + x_d^2}.$$

By doing that, we may see that the problem of concentric circles in $\mathbb{R}^d$ is equivalent to the problem of positive intervals in $\mathbb{R}$ (it is easy to see that ϕ maps points with the same radius to a unique point in $[0, +\infty[$), and consequently we may write that

$$m_{\mathcal{H}}(N) = \frac{N^2}{2} + \frac{N}{2} + 1$$

which is independant of d. As the largest value of N for which $m_{\mathcal{H}}(N) = 2^N$ is 2 ($m_{\mathcal{H}}(3) = 7$), we have that $d_{VC} = 2$.

Problem 2.4

We proceed by constructing a specific set of dichotomies for N points so that among the 2^N possible dichotomies on N points, we select those that contain at most $k - 1$ points labelled (-1). More precisely, we consider the following dichotomies.

- The dichotomies that contain no (-1). We have only $1 = \binom{N}{0}$ of those.
- The dichotomies that contain a unique (-1). We have $N = \binom{N}{1}$ of those.
- The dichotomies that contain exactly two (-1)s. We have $\binom{N}{2}$ of those.
- $\cdots$
- The dichotomies that contain exactly $k - 1$ (-1)s. We have $\binom{N}{k-1}$ of those.

In total, we have exactly $\sum_{i=0}^{k-1} \binom{N}{i}$ such dichotomies. Moreover, these dichotomies do not shatter any subset of k variables because to do that, we would need one dichotomy that contains k (-1)s, which is not the case in our set. So, we may conclude that

$$B(N, k) \geq \sum_{i=0}^{k-1} \binom{N}{i}$$

and with Sauer's lemma, we get

$$B(N, k) = \sum_{i=0}^{k-1} \binom{N}{i}.$$

Problem 2.5

To prove the inequality, we begin with the case $D = 0$. Here, it is easy to see that

$$1 = \binom{N}{0} \leq N^0 + 1 = 2.$$

Now, we assume the result is correct for D ($D \geq 1$), and we will prove it for $D + 1$. We may write that

$$\begin{aligned}\sum_{i=0}^{D+1}\binom{N}{i} &= \sum_{i=0}^{D}\binom{N}{i}+\binom{N}{D+1}\\ &\leq N^D+1+\binom{N}{D+1}\\ &\leq N^D+1+\frac{N!}{(D+1)!(N-D-1)!}.\end{aligned}$$

To continue, we have to prove that

$$\frac{N!}{(N-D-1)!}\leq N^{D+1},$$

which is equivalent to

$$\frac{1}{N^{D+1}}\cdot\frac{N!}{(N-D-1)!}\leq 1.$$

To see this, it suffices to notice that

$$\frac{1}{N^{D+1}}\cdot\frac{N!}{(N-D-1)!}=\frac{1}{N^{D+1}}\prod_{i=0}^{D}(N-i)=\prod_{i=0}^{D}\frac{N-i}{N^{D+1}}\leq 1.$$

So, we are now able to write that

$$\begin{aligned}\sum_{i=0}^{D+1}\binom{N}{i} &\leq N^D+1+\frac{N!}{(D+1)!(N-D-1)!}\\ &\leq N^D+1+\frac{N^{D+1}}{(D+1)!}.\end{aligned}$$

As $D\geq 1$, we have $(D+1)!\geq 2$, and consequently

$$\frac{1}{(D+1)!}\leq\frac{1}{2},$$

which enables us to write that

$$\begin{aligned}\sum_{i=0}^{D+1}\binom{N}{i} &\leq N^D+1+\frac{N^{D+1}}{(D+1)!}\\ &\leq N^D+1+\frac{N^{D+1}}{2}.\end{aligned}$$

Moreover, as we assumed $N\geq D+1$ (if not, we trivially have the result, as in this case $\binom{N}{D+1}=0$), we get $N\geq 2$ and consequently

$$\frac{1}{N}<\frac{1}{2}\Leftrightarrow\frac{N^D}{N^{D+1}}<\frac{1}{2}\Leftrightarrow N^D<\frac{N^{D+1}}{2},$$

which allows us to get our result as we now have

$$\begin{aligned}\sum_{i=0}^{D+1}\binom{N}{i} &\leq N^D+1+\frac{N^{D+1}}{2}\\ &\leq \frac{N^{D+1}}{2}+1+\frac{N^{D+1}}{2}=N^{D+1}+1.\end{aligned}$$

Problem 2.6

As we have $N \geq d$, we may write that $N/d \geq 1$, and also that $(N/d)^{d-i} \geq 1$ for $i = 0, \cdots, d$. Now, we have that

$$\begin{aligned}
\sum_{i=0}^{d} \binom{N}{i} &= \sum_{i=0}^{d} \binom{N}{i} \cdot 1 \\
&\leq \sum_{i=0}^{d} \binom{N}{i} \left(\frac{N}{d}\right)^{d-i} \\
&\leq \left(\frac{N}{d}\right)^{d} \sum_{i=0}^{d} \binom{N}{i} \left(\frac{d}{N}\right)^{i} \\
&\leq \left(\frac{N}{d}\right)^{d} \sum_{i=0}^{N} \binom{N}{i} \left(\frac{d}{N}\right)^{i}.
\end{aligned}$$

Moreover, we also have that

$$\begin{aligned}
\sum_{i=0}^{N} \binom{N}{i} \left(\frac{d}{N}\right)^{i} &= \sum_{i=0}^{N} \binom{N}{i} 1^{N-i} \cdot \left(\frac{d}{N}\right)^{i} \\
&= \left(1 + \frac{d}{N}\right)^{N} \leq e^{d}.
\end{aligned}$$

In conclusion, we have proven that

$$\sum_{i=0}^{d} \binom{N}{i} \leq \left(\frac{N}{d}\right)^{d} \cdot e^{d} = \left(\frac{eN}{d}\right)^{d}.$$

As we already know that

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{d_{VC}} \binom{N}{i},$$

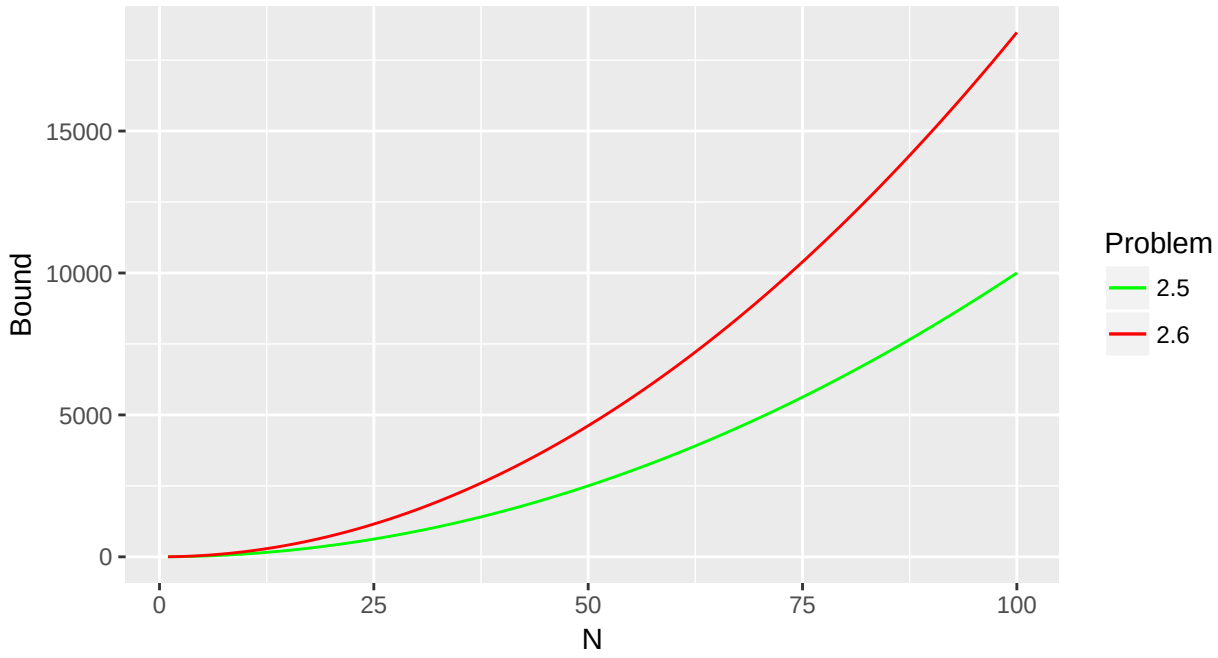
we immediately get that

$$m_{\mathcal{H}}(N) \leq \left(\frac{eN}{d_{VC}}\right)^{d_{VC}}$$

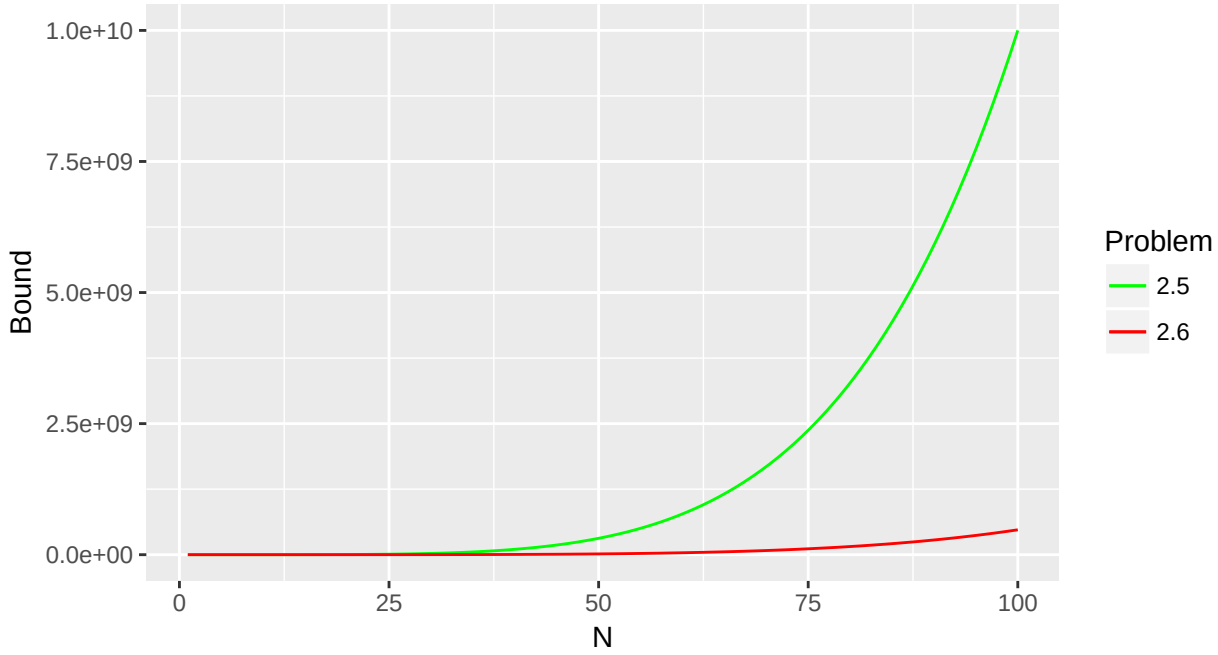
for $N \geq d_{VC}$.

Problem 2.7

We plot below the bounds for $m_{\mathcal{H}}(N)$ given in Problems 2.5 and 2.6 for $d_{VC} = 2$.



Now, we do the same for $d_{VC} = 5$.



For small VC dimensions ($d_{VC} = 2$), it seems that the polynomial bound (Problem 2.5) is better than the exponential one (Problem 2.6); however, for bigger VC dimensions ($d_{VC} = 5$), the exponential bound is way better than the polynomial one.

Problem 2.8

We have only two cases for the growth function : either $d_{VC} = +\infty$ and $m_{\mathcal{H}}(N)$ is equal to 2^N for all N, or d_{VC} is finite and $m_{\mathcal{H}}(N)$ is bounded by $N^{d_{VC}} + 1$.

If $m_{\mathcal{H}}(N) = 1 + N$, we have $d_{VC} = 1$ (as $m_{\mathcal{H}}(2) = 3 < 2^2$). So it must be bounded by $N + 1$ for all N, which

is obviously the case here. In conclusion, $m_{\mathcal{H}}(N) = N + 1$ is a possible growth function.

If $m_{\mathcal{H}}(N) = 1 + N + N(N-1)/2$, we have $d_{VC} = 2$ (as $m_{\mathcal{H}}(3) = 7 < 2^3$). So it must be bounded by $N^2 + 1$ for all N, which is also the case as $N \geq 1$. In conclusion, $m_{\mathcal{H}}(N) = 1 + N + N(N-1)/2$ is a possible growth function.

Obviously $m_{\mathcal{H}}(N) = 2^N$ is a possible growth function (when $d_{VC} = +\infty$).

If $m_{\mathcal{H}}(N) = 2^{\lfloor\sqrt{N}\rfloor}$, we have $d_{VC} = 1$ (as $m_{\mathcal{H}}(2) = 2 < 2^2$). Consequently, it must be bounded by $N + 1$ for all N, which is not true (for $N = 25$ for example). In conclusion, $m_{\mathcal{H}}(N) = 2^{\lfloor\sqrt{N}\rfloor}$ is not a possible growth function.

If $m_{\mathcal{H}}(N) = 2^{\lfloor N/2\rfloor}$, we have $d_{VC} = 0$ (as $m_{\mathcal{H}}(1) = 1 < 2^1$). Consequently, it must be bounded by $N^0 + 1 = 2$ for all N, which is not true (for $N = 4$ for example). In conclusion, $m_{\mathcal{H}}(N) = 2^{\lfloor N/2\rfloor}$ is not a possible growth function.

Problem 2.9

We first define $C(N + 1, d + 1)$ as the number of distinct dichotomies applied to $N + 1$ points that can be generated by an hyperplane hypothesis $\{w \in \mathbb{R}^{d+1} : w^T x = 0\}$. Now, let $x_1, \cdots, x_{N+1} \in \{1\} \times \mathbb{R}^d$ be $N + 1$ points that have no subset of size less than $N + 1$ that is linearly dependent and $(y_1, \cdots, y_N) \in \{-1, 1\}^N$ an hyperplane generated dichotomy on $x_1, \cdots, x_N$; so, there exists w such as

$$(\text{sign}(w^T x_1), \cdots, \text{sign}(w^T x_N)) = (y_1, \cdots, y_N).$$

For the other point x_{N+1}, we also get a dichotomy on $N + 1$ points namely

$$(\text{sign}(w^T x_1), \cdots, \text{sign}(w^T x_{N+1})) = (y_1, \cdots, y_N, \text{sign}(w^T x_{N+1})),$$

thus for every dichotomy over N points there is at least one dichotomy over $N+1$ points. Notice that different dichotomies over N points define different dichotomies over $N + 1$ points, since they differ somewhere in the first N coordinates. Now, potentially, the additional dichotomy

$$(y_1, \cdots, y_N, -\text{sign}(w^T x_{N+1}))$$

is also possible, by some other set of weights. In this way $C(N + 1, d + 1)$ can be higher than $C(N, d + 1)$, so let us write

$$C(N + 1, d + 1) = C(N, d + 1) + D$$

with D the number of additional dichotomies generated by the addition of x_{N+1}.

We will now prove that there exists two hyperplanes that generate the dichotomies $(y_1, \cdots, y_N, +1)$ and $(y_1, \cdots, y_N, -1)$ on $x_1, \cdots, x_{N+1}$ if and only if there exists an hyperplane passing through x_{N+1} that generates the dichotomy $(y_1, \cdots, y_N)$. We begin by showing that the condition is sufficient. We have $w \in \mathbb{R}^{d+1}$ such as $w^T x_{N+1} = 0$ and

$$(\text{sign}(w^T x_1), \cdots, \text{sign}(w^T x_N)) = (y_1, \cdots, y_N),$$

we define $w_1 = w + \epsilon x_{N+1}$ and $w_2 = w - \epsilon x_{N+1}$ with a very small $\epsilon > 0$. Now, it is easy to see that

$$y_k w_1^T x_k = y_k w^T x_k + \epsilon y_k x_{N+1}^T x_k > 0 \text{ and } y_k w_2^T x_k = y_k w^T x_k - \epsilon y_k x_{N+1}^T x_k > 0$$

for $k = 1, \cdots, N$ provided ϵ is sufficiently small, and

$$w_1^T x_{N+1} = w^T x_{N+1} + \epsilon x_{N+1}^T x_{N+1} > 0 \text{ and } w_2^T x_{N+1} = w^T x_{N+1} - \epsilon x_{N+1}^T x_{N+1} < 0.$$

We have now shown that there exist two hyperplanes that generate the dichotomies $(y_1, \cdots, y_N, +1)$ and $(y_1, \cdots, y_N, -1)$ on $x_1, \cdots, x_{N+1}$. Now, we show that the condition is also necessary. We have w_1 and w_2

that generate the dichotomies $(y_1, \cdots, y_N, +1)$ and $(y_1, \cdots, y_N, -1)$ on $x_1, \cdots, x_{N+1}$ respectively, let us now define

$$w = (-w_2^T x_{N+1})w_1 + (w_1^T x_{N+1})w_2.$$

It is easy to see that

$$w^T x_{N+1} = (-w_2^T x_{N+1})(w_1^T x_{N+1}) + (w_1^T x_{N+1})(w_2^T x_{N+1}) = 0,$$

and for $k = 1, \cdots, N$,

$$y_k w^T x_k = (-w_2^T x_{N+1}) y_k (w_1^T x_k) + (w_1^T x_{N+1}) y_k (w_2^T x_k) > 0.$$

We have now shown that there exists an hyperplane passing through x_{N+1} that generates $(y_1, \cdots, y_N)$. We are now able to state that the number of additional dichotomies D generated by the addition of x_{N+1} is equal to the number of hyperplanes passing through x_{N+1} generating the dichotomy $(y_1, \cdots, y_N)$ on $x_1, \cdots, x_N$. As the dimension of an hyperplane $\{w \in \mathbb{R}^{d+1} : w^T x_{N+1} = 0\}$ passing through x_{N+1} is actually the dimension of an hyperplane orthogonal to x_{N+1}, namely d, we have that $D = C(N, d)$. We finally have our recursion formula :

$$C(N+1, d+1) = C(N, d+1) + C(N, d).$$

Intuitively, we have shown that if we assume that one of the weight vectors w that generates $(y_1, \cdots, y_N)$ passes directly through x_{N+1}, it is clear that by slight changes in the angle of the hyperplane we will be able to move the hyperplane slightly to this side or the other of x_{N+1}, thus getting a value of either $+1$ or -1 on it. So in this case, both $(y_1, \cdots, y_N, +1)$ and $(y_1, \cdots, y_N, -1)$ are possible, and there is one additional possible dichotomy beyond $C(N, d+1)$ (that is counted in D). On the other hand, if no hyperplane that passes through x_{N+1} (and generates $(y_1, \cdots, y_N)$ on the first N vectors) exists, then it means that the point lies in one side of all the hyperplanes that generate the old dichotomy, hence we will not be able to achieve both dichotomies, unlike before. We have thus seen that D is the number of those dichotomies over N points that are realized by a hyperplane that passes through a certain fixed point x_{N+1}. Now, by forcing the hyperplane to pass through a certain fixed point, we are in fact moving the problem to one in d dimensions, instead of $d+1$ dimensions.

We are now ready to prove the formula that computes the value of $C(N, d+1)$ for every N and d if we assume that the N points have no subset of size less than N that is linearly dependent,

$$C(N, d+1) = 2 \sum_{i=0}^{d} \binom{N-1}{i}.$$

We will proceed by induction and we begin with case where $N = 1$. We obviously have $C(1, d+1) = 2$ for any d and

$$2 \sum_{i=0}^{d} \binom{0}{i} = 2.$$

We now assume the result is true for N and we will prove it for $N+1$. We have

$$\begin{aligned}
C(N+1, d+1) &= C(N, d+1) + C(N, d) \\
&= 2 \sum_{i=0}^{d} \binom{N-1}{i} + 2 \sum_{i=0}^{d-1} \binom{N-1}{i} \\
&= 2 \sum_{i=0}^{d} \binom{N-1}{i} + 2 \sum_{i=1}^{d} \binom{N-1}{i-1} \\
&= 2 \sum_{i=0}^{d} \left(\binom{N-1}{i} + \binom{N-1}{i-1} \right) \\
&= 2 \sum_{i=0}^{d} \binom{N}{i}.
\end{aligned}$$

Now it remains only to apply the above result to the growth function of the d-dimensional perceptron. We defined the number of dichotomies on N points in $d+1$ dimensions as $C(N, d+1)$ when the N points have no subset of size less than N that is linearly dependent, however it is pretty clear that if this is not the case, we will have far less dichotomies. So, as the growth function is defined as a maximum, we may write that

$$m_{\mathcal{H}}(N) = C(N, d+1) = 2\sum_{i=0}^{d} \binom{N-1}{i}.$$

By using the above formula, we get immediately that

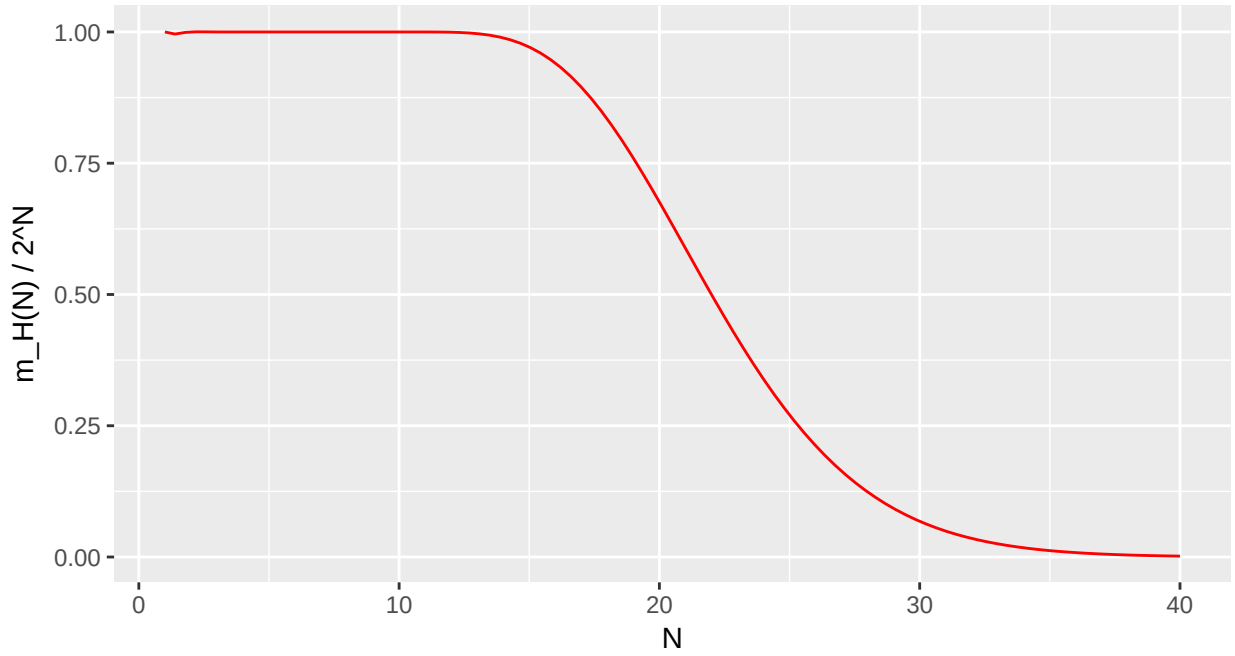
$$m_{\mathcal{H}}(d+1) = 2\left[\binom{d}{0} + \binom{d}{1} + \cdots + \binom{d}{d}\right] = 2 \cdot 2^d = 2^{d+1}$$

and

$$m_{\mathcal{H}}(d+2) = 2\left[\binom{d+1}{0} + \binom{d+1}{1} + \cdots + \binom{d+1}{d}\right] = 2(2^{d+1} - 1) < 2^{d+2},$$

which proves that $d_{VC} = d + 1$ for the d-dimensional perceptron.

Below, we plot $m_{\mathcal{H}}(N)/2^N$ for $d = 10$ and $N \in [1, 40]$.



The number of dichotomies that can be implemented in proportion to the number of possible dichotomies shrinks, since the total number of possible dichotomies is 2^N. For small N, the number of dichotomies is maximal, i.e. 2^N. When N becomes large, the number of possible dichotomies becomes only polynomial and so reach a minimum.

First, we have to remark that a dichotomy is separable if and only if such a dichotomy can be generated by a perceptron. So, it is easy to see that

$$\mathbb{P}(\text{Random dichotomy is separable}) = \frac{\text{Number of separable dichotomies}}{\text{Total number of dichotomies}} \leq \frac{m_{\mathcal{H}}(N)}{2^N}.$$

If we consider $d = 10$, for $N = 10$ we get $\mathbb{P}(\text{Random dichotomy is separable}) = 1$, for $N = 20$ we get $\mathbb{P}(\text{Random dichotomy is separable}) = 0.6761971$, and for $N = 40$ we get $\mathbb{P}(\text{Random dichotomy is separable}) = 0.0016889$.

Problem 2.10

Let us begin with an example : let us say we have 3 ways to dichotomize two points x_1, x_2 ($[1,1]$, $[1,-1]$ and $[-1,1]$) and 2 ways to dichotomize another two points x_3, x_4 ($[1,-1]$ and $[-1,-1]$). So, for each of the 3 ways for the first two points there are at most 2 ways to dichotomize the second two points. In this case, we have at most $3 \times 2 = 6$ ways to dichotomize all four points ($[1,1,1,-1]$, $[1,1,-1,-1]$, $[1,-1,1,-1]$,$[1,-1,-1,-1]$, $[-1,1,1,-1]$, $[-1,1,-1,-1]$).

With this reasoning, let us say that $m_{\mathcal{H}}(N) = p$, now if we partition any set of $2N$ points into two sets of N points each, each of these two partitions will produce p dichotomies at most. If we now combine these two sets, then the maximum number of dichotomies will be the product of p by p. We may conclude that

$$m_{\mathcal{H}}(2N) = m_{\mathcal{H}}(N)^2.$$

If we combine the result above with the VC generalization bound, we get that

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(2N)}{\delta}} \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(N)^2}{\delta}}.$$

Problem 2.11

In the case where $N = 100$, the VC generalization bound tells us that

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{100} \ln \frac{4(2 \cdot 100 + 1)}{0.1}} = E_{in}(g) + 0.8481596$$

with probability at least 90%. When $N = 10000$, we get that

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{10000} \ln \frac{4(2 \cdot 10000 + 1)}{0.1}} = E_{in}(g) + 0.1042782$$

with probability at least 90%.

Problem 2.12

We have the following implicit bound for the sample complexity N (with $d_{VC} = 10$, $\epsilon = 0.05$, and $\delta = 0.05$),

$$N \geq \frac{8}{0.05^2} \ln\Big(\frac{4[(2N)^{10} + 1]}{0.05}\Big).$$

To determine N, we will use an iterative process with an initial guess of $N = 1000$ in the RHS. We get

$$N \geq \frac{8}{0.05^2} \ln\Big(\frac{4[(2 \cdot 1000)^{10} + 1]}{0.05}\Big) \approx 2.57251 \times 10^5.$$

We then try the new value $N = 2.57251 \times 10^5$ in the RHS and iterate this process, rapidly converging to an estimate of $N \approx 4.52957 \times 10^5$.

Problem 2.13

(a) Let $d_{VC}(\mathcal{H}) = d$, by definition we have that $m_{\mathcal{H}}(d) = 2^d$. So, we may write that

$$\begin{aligned} 2^d = m_{\mathcal{H}}(d) &= \max_{x_1,\cdots,x_d} |\mathcal{H}(x_1,\cdots,x_d)| \\ &= \max_{x_1,\cdots,x_d} |\{(h(x_1),\cdots,h(x_d)) : h \in \mathcal{H}\}| \\ &\leq |\mathcal{H}| = M. \end{aligned}$$

In conclusion we have $d \leq \log_2(M)$.

(b) At worst, we have that $\cap_{k=1}^K \mathcal{H}_k = \{h\}$, in this case its VC dimension is trivially 0 as $m_{\mathcal{H}}(N) = 1$ for all N. So, we have that $d_{VC}(\cap_{k=1}^K \mathcal{H}_k) \geq 0$.

Now, we will prove that

$$d_{VC}(\cap_{k=1}^K \mathcal{H}_k) \leq \min_{1\leq k\leq K} d_{VC}(\mathcal{H}_k).$$

To do that we assume that

$$d_{VC}(\cap_{k=1}^K \mathcal{H}_k) > \min_{1\leq k\leq K} d_{VC}(\mathcal{H}_k) = d,$$

which means that $\cap_{k=1}^K \mathcal{H}_k$ can shatter $d+1$ points, let $x_1,\cdots,x_{d+1}$ be these points. We may write that

$$\begin{aligned} \{-1,+1\}^{d+1} = \cap_{k=1}^K \mathcal{H}_k(x_1,\cdots,x_{d+1}) &= \{(h(x_1),\cdots,h(x_{d+1})) : h \in \cap_{k=1}^K \mathcal{H}_k\} \\ &\subset \{(h(x_1),\cdots,h(x_{d+1})) : h \in \mathcal{H}_k\} = \mathcal{H}_k(x_1,\cdots,x_{d+1}) \end{aligned}$$

for all $k = 1,\cdots,K$. If we compute the cardinality of these sets, we get the following inequality

$$2^{d+1} \leq |\{(h(x_1),\cdots,h(x_{d+1})) : h \in \mathcal{H}_k\}| \leq 2^{d+1},$$

which means that

$$|\{(h(x_1),\cdots,h(x_{d+1})) : h \in \mathcal{H}_k\}| = 2^{d+1}$$

for all $k = 1,\cdots,K$. So, any $\mathcal{H}_k$ can shatter $d+1$ points, if we let $\min_{1\leq k\leq K} d_{VC}(\mathcal{H}_k) = d_{VC}(\mathcal{H}_{k_0})$, we get that

$$d = d_{VC}(\mathcal{H}_{k_0}) \geq d+1$$

which is not possible. In conclusion, we have

$$0 \leq d_{VC}(\cap_{k=1}^K \mathcal{H}_k) \leq \min_{1\leq k\leq K} d_{VC}(\mathcal{H}_k).$$

(c) Let $d_{VC}(\mathcal{H}_k) = d_k$ for all $k = 1,\cdots,K$. This means that $\mathcal{H}_k$ shatters d_k points $x_1,\cdots,x_{d_k}$, or in other words

$$\mathcal{H}_k(x_1,\cdots,x_{d_k}) = \{-1,+1\}^{d_k}$$

for all $k = 1,\cdots,K$. It is easy to see that

$$\{-1,+1\}^{d_k} = \{(h(x_1),\cdots,h(x_{d_k})) : h \in \mathcal{H}_k\} \subset \{(h(x_1),\cdots,h(x_{d_k})) : h \in \cup_{k=1}^K \mathcal{H}_k\},$$

if we compute the cardinality of these sets, we get

$$2^{d_k} \leq |\{(h(x_1),\cdots,h(x_{d_k})) : h \in \cup_{k=1}^K \mathcal{H}_k\}| \leq 2^{d_k}.$$

So, here again we conclude that

$$|\{(h(x_1),\cdots,h(x_{d_k})) : h \in \cup_{k=1}^K \mathcal{H}_k\}| = 2^{d_k},$$

or more simply that $m_{\cup_{k=1}^K \mathcal{H}_k}(d_k) = 2^{d_k}$ for all $k = 1, \cdots, K$. In conclusion, we have that $d_{VC}(\cup_{k=1}^K \mathcal{H}_k) \geq d_k$ for each k, and consequently

$$d_{VC}(\cup_{k=1}^K \mathcal{H}_k) \geq \max_{1 \leq k \leq K} d_k = \max_{1 \leq k \leq K} d_{VC}(\mathcal{H}_k).$$

Let us derive the upper bound for $K = 2$ first, to do that we let $d_{VC}(\mathcal{H}_1) = d_1$ and $d_{VC}(\mathcal{H}_2) = d_2$. It is pretty clear that the number of dichotomies generated by $\mathcal{H}_1 \cup \mathcal{H}_2$ is at most the sum of the dichotomies generated by $\mathcal{H}_1$ and by $\mathcal{H}_2$. We are now able to write that

$$\begin{aligned}
m_{\mathcal{H}_1 \cup \mathcal{H}_2}(N) &\leq m_{\mathcal{H}_1}(N) + m_{\mathcal{H}_2}(N) \\
&\leq \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=0}^{d_2} \binom{N}{i} \\
&\leq \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=0}^{d_2} \binom{N}{N-i} \\
&\leq \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=N-d_2}^{N} \binom{N}{i} \\
&< \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=d_1+1}^{N-d_2-1} \binom{N}{i} + \sum_{i=N-d_2}^{N} \binom{N}{i} = \sum_{i=0}^{N} \binom{N}{i} = 2^N
\end{aligned}$$

for all N such as $d_1 + 1 \leq N - d_2 - 1 \Leftrightarrow N \geq d_1 + d_2 + 2$. We can now deduce that

$$d_{VC}(\mathcal{H}_1 \cup \mathcal{H}_2) \leq d_1 + d_2 + 1.$$

We will now prove by induction the general upper bound

$$d_{VC}(\cup_{k=1}^K \mathcal{H}_k) \leq K - 1 + \sum_{k=1}^{K} d_{vc}(\mathcal{H}_k).$$

For $K = 2$, we immediately have

$$d_{VC}(\mathcal{H}_1 \cup \mathcal{H}_2) \leq 1 + \sum_{k=1}^{2} d_{VC}(\mathcal{H}_k)$$

from the upper bound above. Now, if we assume the result is correct for $K - 1$, we will prove it for K. We have that

$$\begin{aligned}
d_{VC}(\cup_{k=1}^K \mathcal{H}_k) &= d_{VC}((\cup_{k=1}^{K-1} \mathcal{H}_k) \cup \mathcal{H}_K) \\
&\leq 1 + d_{VC}(\cup_{k=1}^{K-1} \mathcal{H}_k) + d_{VC}(\mathcal{H}_K) \\
&\leq 1 + (K-2) + \sum_{k=1}^{K-1} d_{VC}(\mathcal{H}_k) + d_{VC}(\mathcal{H}_K) \\
&\leq K - 1 + \sum_{k=1}^{K} d_{VC}(\mathcal{H}_k).
\end{aligned}$$

Problem 2.14

(a) By problem 2.13, we may write that

$$\begin{aligned} d_{VC}(\mathcal{H}) &\leq K - 1 + \sum_{k=1}^{K} d_{VC} \\ &\leq K - 1 + K d_{VC} \\ &< K(d_{VC} + 1). \end{aligned}$$

(b) We already know that

$$m_{\mathcal{H}_k}(l) \leq l^{d_{VC}} + 1$$

for every $k = 1, \cdots, K$, which means that every $\mathcal{H}_k$ can only generate $l^{d_{VC}} + 1$ dichotomies at most. Let us take a look at what happens when we consider the number of dichotomies generated by $\mathcal{H}$. It is pretty obvious that for any l points $x_1, \cdots, x_l$, we have

$$\{(h(x_1), \cdots, h(x_l)) : h \in \mathcal{H}\} \subset \cup_{k=1}^{K} \{(h(x_1), \cdots, h(x_l)) : h \in \mathcal{H}_k\},$$

if we compute the cardinality, we get that

$$|\{(h(x_1), \cdots, h(x_l)) : h \in \mathcal{H}\}| \leq \sum_{k=1}^{K} |\{(h(x_1), \cdots, h(x_l)) : h \in \mathcal{H}_k\}|.$$

So, we get that

$$m_{\mathcal{H}}(l) \leq \sum_{k=1}^{K} m_{\mathcal{H}_k}(l) \leq K(l^{d_{VC}} + 1) \leq 2K l^{d_{VC}},$$

as

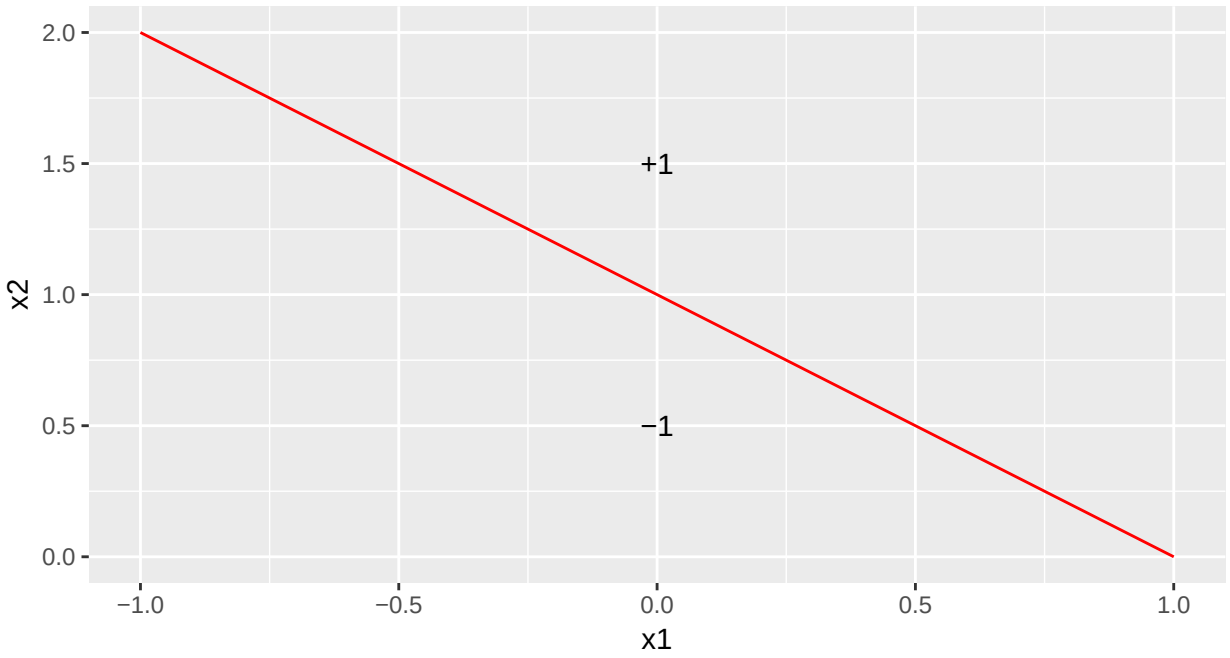
$$\begin{aligned} &\Leftrightarrow \quad K(l^{d_{VC}} + 1) \leq 2K l^{d_{VC}} \\ &\Leftrightarrow \quad 1 \leq l^{d_{VC}} \end{aligned}$$

which is true if we assume $l \geq 1$. By hypothesis, we now have that $m_{\mathcal{H}}(l) < 2^l$, and consequently $d_{VC}(\mathcal{H}) < l$, which obviously implies that $d_{VC}(\mathcal{H}) \leq l$.

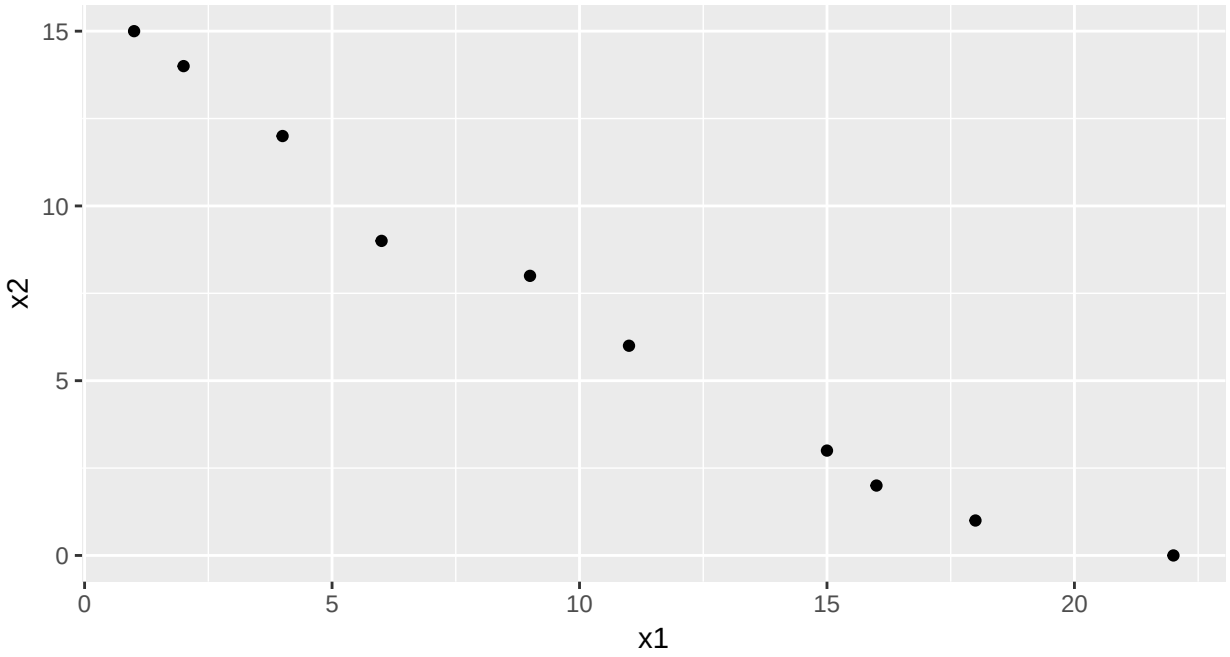
(c)

Problem 2.15

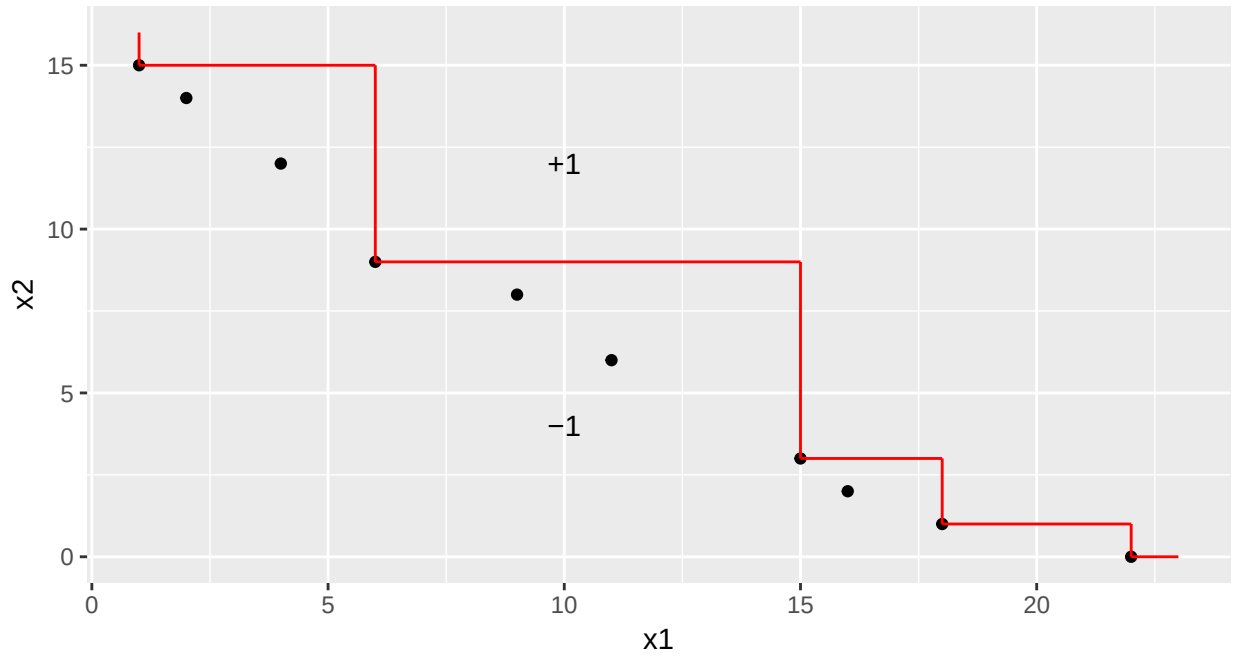
(a) We plot below a monotonic classifier in two dimensions.



(*b*) To compute $m_{\mathcal{H}}(N)$, we consider the N points generated by first choosing one point, then generating the next point by increasing the first component and decreasing the second component until N points are obtained. Below, we gave an example of this process for 10 points.



Then, we consider a random dichotomy on these points $(+1, -1, -1, +1, -1, -1, +1, -1, +1, +1)$. It is now easy to see that the piecewise linear function plotted below (which passes through each $+1$ labelled point and is above any -1 labelled point) actually characterizes a monotonically increasing hypothesis and generates our dichotomy.



In conclusion, our hypothesis set $\mathcal{H}$ is able to generate any dichotomy on N points, so it is obvious that $m_{\mathcal{H}}(N) = 2^N$ for all N. Consequently, we get that $d_{VC}(\mathcal{H}) = \infty$.

Problem 2.16

(a) We begin by choosing $D+1$ distinct points in $\mathbb{R}$: $x_0, x_1, \cdots, x_D$. We let X be the following Vandermonde matrix

$$X = \begin{pmatrix} 1 & x_0^1 & \cdots & x_0^D \\ 1 & x_1^1 & \cdots & x_1^D \\ \vdots & \vdots & & \vdots \\ 1 & x_D^1 & \cdots & x_D^D \end{pmatrix}.$$

We see immediately that X is a square matrix with $D+1$ dimensions, and that $\det X \neq 0$ as the x_k are all different. Then we consider any dichotomy $y = (y_0, \cdots, y_D)^T \in \{-1, +1\}^{D+1}$, and we let $c = (c_0, \cdots, c_D)^T = X^{-1}y$. We now have immediately that $Xc = y$, and consequently

$$h_c(x_k) = \text{sign}(\sum_{i=0}^{D} c_i x_k^i) = y_k$$

for all $k = 0, \cdots, D$. In conclusion, our hypothesis set $\mathcal{H}$ shatters $x_0, \cdots, x_D$, so $m_{\mathcal{H}}(D+1) = 2^{D+1}$ and $d_{VC}(\mathcal{H}) \geq D+1$.

(b) In this case, we consider any $D+2$ points in $\mathbb{R}$: $x_0, x_1, \cdots, x_{D+1}$. Then the $D+2$ vectors $(x_k^0, x_k^1, \cdots, x_k^D)$ $(k = 0, \cdots, D+1)$ are $D+2$ vectors in $D+1$ dimensions, so we may conclude that they are linearly dependant. Consequently, there exists an index l and $D+1$ coefficients a_k (not all equal to zero) so that

$$(x_l^0, x_l^1, \cdots, x_l^D) = \sum_{k \neq l} a_k (x_k^0, x_k^1, \cdots, x_k^D).$$

Now, we choose a specific dichotomy y defined by $y_k = \text{sign}(a_k)$ if $a_k \neq 0$ (y_k can be $+1$ or -1 if $a_k = 0$), and $y_l = -1$. Then we consider any $c = (c_0, \cdots, c_D)^T$ and we get that

$$(x_l^0, x_l^1, \cdots, x_l^D) \begin{pmatrix} c_0 \\ \vdots \\ c_D \end{pmatrix} = \sum_{k \neq l} a_k (x_k^0, x_k^1, \cdots, x_k^D) \begin{pmatrix} c_0 \\ \vdots \\ c_D \end{pmatrix} = \sum_{k \neq l} \sum_{i=0}^{D} c_i a_k x_k^i.$$

So, let us assume that there exists $c \in \mathbb{R}^{D+1}$ so that

$$y_k = h_c(x_k) = \text{sign}(\sum_{i=0}^{D} c_i x_k^i)$$

for any k so that $a_k \neq 0$ and

$$y_l = h_c(x_l) = \text{sign}(\sum_{i=0}^{D} c_i x_l^i),$$

in this case we would get

$$\text{sign}(a_k) = y_k = \text{sign}(\sum_{i=0}^{D} c_i x_k^i),$$

which is equivalent to

$$\sum_{i=0}^{D} c_i a_k x_k^i > 0$$

for any k so that $a_k \neq 0$. So, we also get that

$$\sum_{k \neq l} \sum_{i=0}^{D} c_i a_k x_k^i = (x_l^0, x_l^1, \cdots, x_l^D) \begin{pmatrix} c_0 \\ \vdots \\ c_D \end{pmatrix} = \sum_{i=0}^{D} c_i x_l^i > 0,$$

and then

$$y_l = \text{sign}(\sum_{i=0}^{D} c_i x_l^i) = +1.$$

However, we had that $y_l = -1$. So we have a dichotomy $y \in \mathbb{R}^{D+2}$ that cannot be generated by $\mathcal{H}$, this implies that $m_{\mathcal{H}}(D+2) < 2^{D+2}$ and that $d_{VC}(\mathcal{H}) \leq D+1$.

Problem 2.17

Let d_1 and d_2 be the VC dimensions of $\mathcal{H}$ with respect to $\mathcal{X}_1$ and $\mathcal{X}_2$ respectively. By definition, we have that

$$m_{\mathcal{H}}^{\mathcal{X}_2}(d_2+1) < 2^{d_2+1},$$

so if we consider any d_2+1 points $x_1, \cdots, x_{d_2+1} \in \mathcal{X}_1 \subset \mathcal{X}_2$, there exists a dichotomy $y = (y_1, \cdots, y_{d_2+1})^T \in \{-1,+1\}^{d_2+1}$ that cannot be generated by $\mathcal{H}$ on $x_1, \cdots, x_{d_2+1}$. This means that $\mathcal{H}$ does not shatter $x_1, \cdots, x_{d_2+1} \in \mathcal{X}_1$, so we may write that

$$m_{\mathcal{H}}^{\mathcal{X}_1}(d_2+1) < 2^{d_2+1},$$

which implies that $d_1 \leq d_2$.

Let $\mathcal{X}_1 = \{(x^0, x^1, \cdots, x^D) : x \in \mathbb{R}\}$, $\mathcal{X}_2 = \{1\} \times \mathbb{R}^d$, and $\mathcal{H} = \{h : h(x) = \text{sign}(\sum_{i=0}^{D} w_i x_i)\}$ be the hypothesis set of the perceptron in D dimensions. We obviously have that $\mathcal{X}_1 \subset \mathcal{X}_2$, so the result above allows us to write that $d_{VC}^{\mathcal{X}_1}(\mathcal{H}) \leq d_{VC}^{\mathcal{X}_2}(\mathcal{H})$. However, $d_{VC}^{\mathcal{X}_2}(\mathcal{H})$ is actually the VC dimension of the perceptron in D dimensions which we know to be equal to $D+1$, moreover $d_{VC}^{\mathcal{X}_1}(\mathcal{H})$ is also the VC dimension of the hypothesis set introduced in Problem 2.16. Consequently, $d_{VC}^{\mathcal{X}_1}(\mathcal{H}) < D+2$ which means that no $(D+2)$ points are shattered by the Problem 2.16 hypothesis set.

Problem 2.18

First, we choose N points $x_1 = 10^1, x_2 = 10^2, \cdots, x_N = 10^N$ in $\mathbb{R}$, then we let $y = (y_1, \cdots, y_N)^T \in \{-1, +1\}^N$ be any dichotomy. Now, we consider $\alpha = 0.d_1 d_2 \cdots d_N$ with the digit $d_i = 1$ if $y_i = -1$ and $d_i = 2$ if $y_i = +1$; then we immediately have that

$$h_\alpha(x_k) = (-1)^{\lfloor \alpha \cdot 10^k \rfloor} = y_k$$

for all $k = 1 \cdots, N$. We may now conclude that $\mathcal{H}(x_1, \cdots, x_N) = \{-1, +1\}^N$ (or $m_\mathcal{H}(N) = 2^N$) for all N, and so $d_{VC}(\mathcal{H}) = \infty$.

Problem 2.19

(a) Let $x_1, \cdots, x_N$ be N points in $\mathbb{R}^d$. Now, if we fix $h_1 \in \mathcal{H}_1, \cdots, h_K \in \mathcal{H}_K$, with these functions we are able to generate $z_k = (h_1(x_k), \cdots, h_K(x_k)) \in \{-1, +1\}^K$ for $k = 1, \cdots, N$. By definition, $\tilde{\mathcal{H}}$ can only generate $m_{\tilde{\mathcal{H}}}(N)$ dichotomies on $z_1, \cdots, z_N$ at most. This implies that $\tilde{\mathcal{H}}$ can only generate $m_{\tilde{\mathcal{H}}}(N)$ dichotomies on $x_1, \cdots, x_N$ for every choice of $(h_1, \cdots, h_K)$. However, even if we do not know the precise number of hypothesis in each $\mathcal{H}_i$, we can say that, when we consider $x_1, \cdots, x_N$, the effective number of hypothesis in $\mathcal{H}_i$ is actually $m_{\mathcal{H}_i}(N)$ at most. In conclusion, as we have at most $m_{\tilde{\mathcal{H}}}(N)$ dichotomies on $x_1, \cdots, x_N$ for every choice of $(h_1, \cdots, h_K)$, and as we have at most $\prod_{i=1}^K m_{\mathcal{H}_i}(N)$ choices of $(h_1, \cdots, h_k)$ to consider for our N points, we may write that

$$m_\mathcal{H}(N) \leq m_{\tilde{\mathcal{H}}}(N) \prod_{i=1}^K m_{\mathcal{H}_i}(N).$$

(b) We may write that

$$m_{\tilde{\mathcal{H}}}(N) \leq \left(\frac{eN}{\tilde{d}} \right)^{\tilde{d}}$$

and

$$m_{\mathcal{H}_i}(N) \leq \left(\frac{eN}{d_i} \right)^{d_i}$$

for $i = 1, \cdots, K$. Consequently, we have that

$$m_\mathcal{H}(N) \leq \left(\frac{eN}{\tilde{d}} \right)^{\tilde{d}} \prod_{i=1}^K \left(\frac{eN}{d_i} \right)^{d_i} = \frac{(eN)^{\tilde{d} + \sum_{i=1}^K d_i}}{\tilde{d}^{\tilde{d}} \prod_{i=1}^K d_i^{d_i}}.$$

(c) It is easy to notice that

$$\begin{aligned} m_\mathcal{H}(2D \log_2 D) &\leq \frac{(D \cdot 2e \log_2 D)^D}{\tilde{d}^{\tilde{d}} \prod_{i=1}^K d_i^{d_i}} \\ &\leq (D \cdot 2e \log_2 D)^D \\ &< D^{2D} = 2^{2D \log_2 D} \end{aligned}$$

as $\tilde{d}^{\tilde{d}} \prod_{i=1}^K d_i^{d_i} \geq 1$ and $2e \log_2 D < D$. We are now able to conclude that

$$d_{VC}(\mathcal{H}) < 2D \log_2 D.$$

(d) If $\mathcal{H}_i$ and $\tilde{\mathcal{H}}$ are all perceptron hypothesis sets, we get that $d_i = d$ and $\tilde{d} = K$, and $D = Kd + K$. For K and d sufficiently large, we may write that

$$\begin{aligned} d_{VC}(\mathcal{H}) &\leq 2(dK+K)\log_2(dK+K) \\ &\leq 2(2dK)\log_2(2dK) \\ &\leq \frac{4}{\log(2)} dK(\log(dK)+\log(2)) \\ &\leq \frac{8}{\log(2)} dK\log(dK), \end{aligned}$$

which means that

$$d_{VC}(\mathcal{H}) = O(dK\log(dK)).$$

Problem 2.20

To plot the various bounds for $d_{VC} = 50$ and $\delta = 0.05$, we will use the following bound on the growth function

$$m_{\mathcal{H}}(N) \leq N^{d_{VC}} + 1.$$

As the Parrondo and the Devroye bounds are implicit in ϵ, we will first transform these expressions into explicit ones. The Parrondo and Van den Broek bound can be written as

$$\begin{aligned} & \epsilon^2 \leq \frac{1}{N}\left(2\epsilon + \ln\frac{6m_{\mathcal{H}}(2N)}{\delta}\right) \\ \Leftrightarrow \quad & N\epsilon^2 - 2\epsilon - \ln\frac{6m_{\mathcal{H}}(2N)}{\delta} \leq 0. \end{aligned}$$

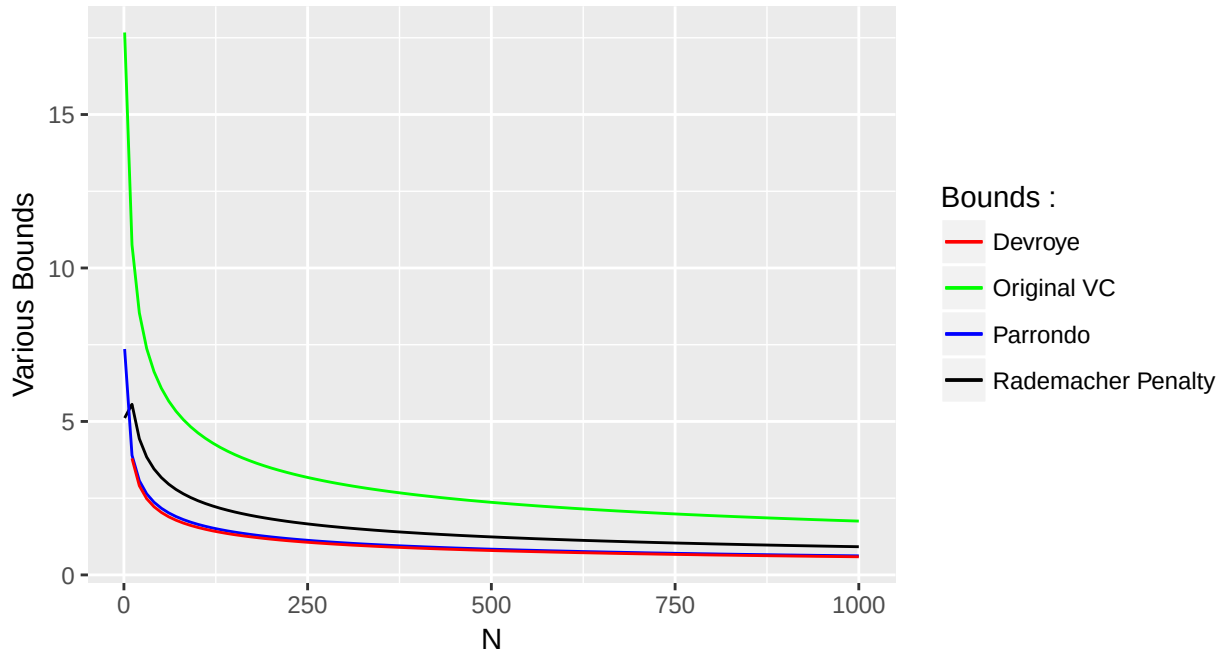
By studying the sign of the polynomial above, we get the following explicit expression for ϵ,

$$\epsilon \leq \frac{1+\sqrt{N\ln\frac{6m_{\mathcal{H}}(2N)}{\delta}+1}}{N}.$$

If we proceed in the same fashion for the Devroye bound, we get the following expression for ϵ,

$$\epsilon \leq \frac{2+\sqrt{(2N-4)\ln\frac{4m_{\mathcal{H}}(N^2)}{\delta}+4}}{2(N-2)}.$$

These are the actual expressions that we will use to plot the various bounds below.



As we may notice on the plot above, the best bounds are the Parrondo and Van den Broek one, and the Devroye one. A more detailed analysis shows that the Parrondo and Van den Broek bound is better when N is small (for example when $N = 5$, we get 5.101362 for the Parrondo and Van den Broek bound, and 5.5931255 for the Devroye bound), and the other way around when N is large (for example when $N = 1000$, we get 0.6213496 for the Parrondo and Van den Broek bound, and 0.5911514 for the Devroye bound).

Problem 2.21

We have, for any $\epsilon > 0$, that

$$\begin{aligned} & \mathbb{P}\left[\frac{E_{out}(g) - E_{in}(g)}{\sqrt{E_{out}(g)}} > \epsilon\right] \leq c m_{\mathcal{H}}(2N) e^{-\frac{\epsilon^2 N}{4}} \\ \Leftrightarrow \quad & \mathbb{P}\left[\frac{E_{out}(g) - E_{in}(g)}{\sqrt{E_{out}(g)}} \leq \epsilon\right] \geq 1 - c m_{\mathcal{H}}(2N) e^{-\frac{\epsilon^2 N}{4}}. \end{aligned}$$

However, if we assume $E_{out}(g) \geq E_{in}(g)$ then a simple sign study shows that

$$\begin{aligned} \frac{E_{out}(g) - E_{in}(g)}{\sqrt{E_{out}(g)}} \leq \epsilon \quad & \Rightarrow \quad \frac{(E_{out}(g) - E_{in}(g))^2}{E_{out}(g)} \leq \epsilon^2 \\ & \Rightarrow \quad E_{out}(g)^2 - (2E_{in}(g) + \epsilon^2) E_{out}(g) + E_{in}(g)^2 \leq 0 \\ & \Rightarrow \quad E_{out}(g) \leq E_{in}(g) + \frac{\epsilon^2}{2}\left(1 + \sqrt{1 + \frac{4}{\epsilon^2} E_{in}(g)}\right). \end{aligned}$$

We are now able to state that

$$\mathbb{P}\left[E_{out}(g) \leq E_{in}(g) + \frac{\epsilon^2}{2}\left(1 + \sqrt{1 + \frac{4}{\epsilon^2} E_{in}(g)}\right)\right] \geq \mathbb{P}\left[\frac{E_{out}(g) - E_{in}(g)}{\sqrt{E_{out}(g)}} \leq \epsilon\right] \geq 1 - c m_{\mathcal{H}}(2N) e^{-\frac{\epsilon^2 N}{4}}$$

for all $\epsilon > 0$. Now, it remains to let

$$\epsilon^2 = \xi = \frac{4}{N} \ln \frac{c m_{\mathcal{H}}(2N)}{\delta},$$

and we may conclude that

$$\mathbb{P}\left[E_{out}(g) \leq E_{in}(g) + \frac{\xi}{2}\left(1 + \sqrt{1 + \frac{4}{\xi} E_{in}(g)}\right)\right] \geq 1 - \delta$$

for all $\delta > 0$.

Problem 2.22

We may write that

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[E_{out}(g^{(\mathcal{D})})] &= \mathbb{E}_{\mathcal{D}}[\mathbb{E}_{\mathbf{x},y}[(g^{(\mathcal{D})}(\mathbf{x}) - y(\mathbf{x}))^2]] \\
&= \mathbb{E}_{\mathbf{x},y}[\mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\mathbf{x}) - y(\mathbf{x}))^2]] \\
&= \mathbb{E}_{\mathbf{x},y}[\mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\mathbf{x})^2] - 2\overline{g}(\mathbf{x})y(\mathbf{x}) + y(\mathbf{x})^2]
\end{aligned}$$

where we used Fubini's theorem to switch between the two expectations. Now, we have that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\mathbf{x})^2] - 2\overline{g}(\mathbf{x})y(\mathbf{x}) + y(\mathbf{x})^2 \\
&= \mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\mathbf{x})^2] - 2\overline{g}(\mathbf{x})(f(\mathbf{x}) + \epsilon) + (f(\mathbf{x}) + \epsilon)^2 \\
&= (\mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\mathbf{x})^2] - \overline{g}(\mathbf{x})^2) + (\overline{g}(\mathbf{x})^2 - 2\overline{g}(\mathbf{x}) + f(\mathbf{x})^2) + \epsilon^2 - 2(\overline{g}(\mathbf{x}) - f(\mathbf{x}))\epsilon \\
&= (\mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\mathbf{x})^2] - 2\mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\mathbf{x})]\overline{g}(\mathbf{x}) + \overline{g}(\mathbf{x})^2) + (\overline{g}(\mathbf{x}) - f(\mathbf{x}))^2 + \epsilon^2 - 2(\overline{g}(\mathbf{x}) - f(\mathbf{x}))\epsilon \\
&= \mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\mathbf{x})^2 - 2g^{(\mathcal{D})}(\mathbf{x})\overline{g}(\mathbf{x}) + \overline{g}(\mathbf{x})^2] + (\overline{g}(\mathbf{x}) - f(\mathbf{x}))^2 + \epsilon^2 - 2(\overline{g}(\mathbf{x}) - f(\mathbf{x}))\epsilon \\
&= \mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\mathbf{x}) - \overline{g}(\mathbf{x}))^2] + (\overline{g}(\mathbf{x}) - f(\mathbf{x}))^2 + \epsilon^2 - 2(\overline{g}(\mathbf{x}) - f(\mathbf{x}))\epsilon \\
&= \mathrm{var}(\mathbf{x}) + \mathrm{bias}(\mathbf{x}) + \epsilon^2 - 2(\overline{g}(\mathbf{x}) - f(\mathbf{x}))\epsilon.
\end{aligned}$$

Now, we have to take the expectation relative to $(\mathbf{x}, y)$ to get our result, we have that

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[E_{out}(g^{(\mathcal{D})})] &= \mathbb{E}_{\mathbf{x},y}[\mathrm{var}(\mathbf{x})] + \mathbb{E}_{\mathbf{x},y}[\mathrm{bias}(\mathbf{x})] + \mathbb{E}_{\mathbf{x},y}[\epsilon^2] - 2\mathbb{E}_{\mathbf{x},y}[(\overline{g}(\mathbf{x}) - f(\mathbf{x}))\epsilon] \\
&= \mathbb{E}_{\mathbf{x}}[\mathrm{var}(\mathbf{x})] + \mathbb{E}_{\mathbf{x}}[\mathrm{bias}(\mathbf{x})] + \mathbb{E}_{y}[\epsilon^2] - 2\mathbb{E}_{\mathbf{x},y}[(\overline{g}(\mathbf{x}) - f(\mathbf{x}))\epsilon] \\
&= \mathrm{var} + \mathrm{bias} + \mathbb{E}_y[(\epsilon - \mathbb{E}_y[\epsilon])^2] - 2\mathbb{E}_{\mathbf{x}}[\mathbb{E}_{y|\mathbf{x}}[(\overline{g}(\mathbf{x}) - f(\mathbf{x}))\epsilon|\mathbf{x}]] \\
&= \mathrm{var} + \mathrm{bias} + \mathrm{var}_y(\epsilon) - 2\mathbb{E}_{\mathbf{x}}[(\overline{g}(\mathbf{x}) - f(\mathbf{x}))\mathbb{E}_{y|\mathbf{x}}[\epsilon|\mathbf{x}]] \\
&= \mathrm{var} + \mathrm{bias} + \sigma^2 - 2\mathbb{E}_{\mathbf{x}}[(\overline{g}(\mathbf{x}) - f(\mathbf{x}))\mathbb{E}_y[\epsilon]] \\
&= \mathrm{var} + \mathrm{bias} + \sigma^2
\end{aligned}$$

as $\mathbb{E}_y[\epsilon] = 0$.

Problem 2.23

(a) Here, we consider the learning model of all hypotheses of the form $h(x) = ax + b$. First, to find the best hypothesis that approximates f, we use the training set $\mathcal{D} = \{(x_1, y_1), (x_2, y_2)\}$ to minimize the in-sample mean squared error

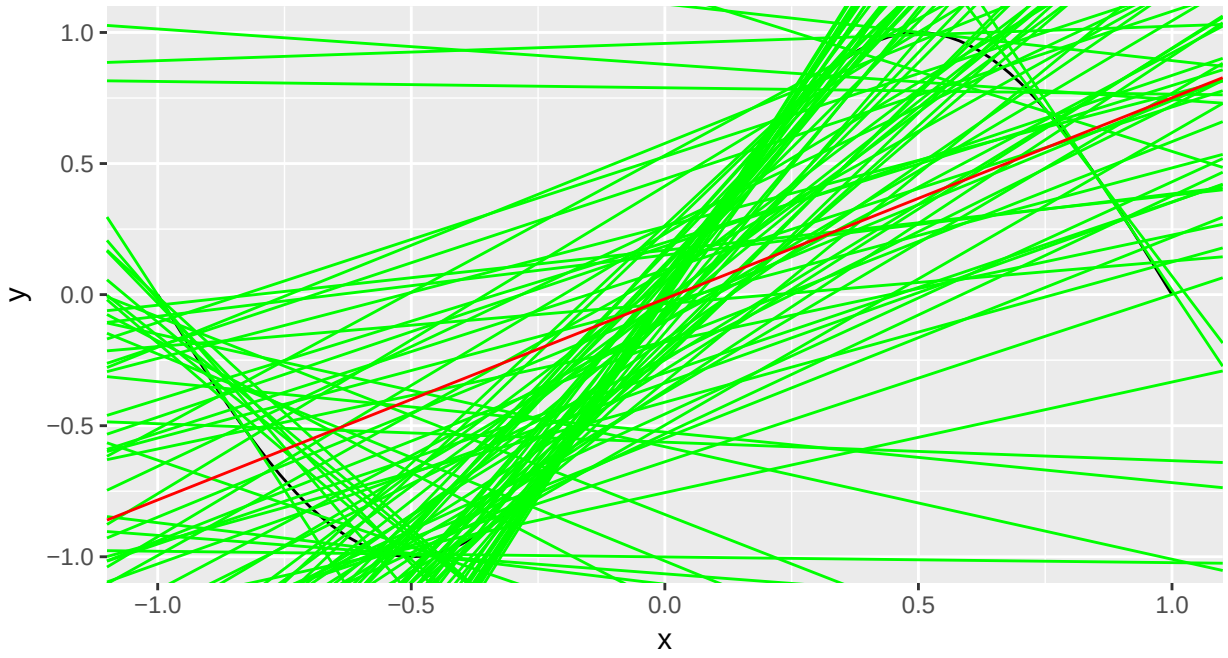
$$E_{in}(h) = \frac{1}{2}\sum_{i=1}^{2}(ax_i + b - y_i)^2$$

with respect to a and b. A simple computation shows that the best hypothesis is given by

$$g(x) = \frac{y_1 - y_2}{x_1 - x_2} x + \frac{x_1 y_2 - x_2 y_1}{x_1 - x_2},$$

which is the line that passes through the points (x_1, y_1) and (x_2, y_2). To compute $\overline{g}$, we will generate many training sets and evaluate a and b for each one of these before averaging them on these training sets to obtain $\overline{a}$ and $\overline{b}$ and consequently $\overline{g}(x) = \overline{a}x + \overline{b}$.

```
f <- function(x)
  sin(pi * x)
plot <- ggplot(data.frame(x = seq(-1, 1, 0.001)), aes(x = x)) +
  stat_function(fun = f, geom = "line")

set.seed(1975)
a.hat <- numeric()
b.hat <- numeric()
for (i in 1:10000) {
  rand <- runif(2, min = -1, max = 1)
  data <- data.frame(x = rand, y = f(rand))
  a <- ((data$y[1] - data$y[2]) / (data$x[1] - data$x[2]))
  a.hat <- c(a.hat, a)
  b <- (data$y[2] * data$x[1] - data$y[1] * data$x[2]) / (data$x[1] - data$x[2])
  b.hat <- c(b.hat, b)
  if (i %% 100 == 0)
    plot <- plot + geom_abline(intercept = b, slope = a, col = "green")
}
a.bar <- mean(a.hat)
b.bar <- mean(b.hat)
plot + geom_abline(intercept = b.bar, slope = a.bar, col = "red")
```



Now, we are able to compute the expected out-of-sample error and its bias and var components by generating a test set and averaging x on this new set.

```
x.new <- runif(100000, min = -1, max = 1)
bias.x <- (a.bar * x.new - b.bar - f(x.new))^2
bias <- round(mean(bias.x), 2)
var.x <- mean((a.hat - a.bar)^2) * x.new^2 +
  2 * mean((a.hat - a.bar) * (b.hat - b.bar)) * x.new +
  mean((b.hat - b.bar)^2)
var <- round(mean(var.x), 2)
```

So, here we get a bias of 0.21 and a variance of 1.69, which gives us an expected out-of sample error of 1.9.

(*b*) Now, we consider the learning model of all hypotheses of the form $h(x) = ax$. Here again, to find the best hypothesis that approximates f, we use the training set $\mathcal{D} = \{(x_1, y_1), (x_2, y_2)\}$ to minimize the in-sample mean squared error

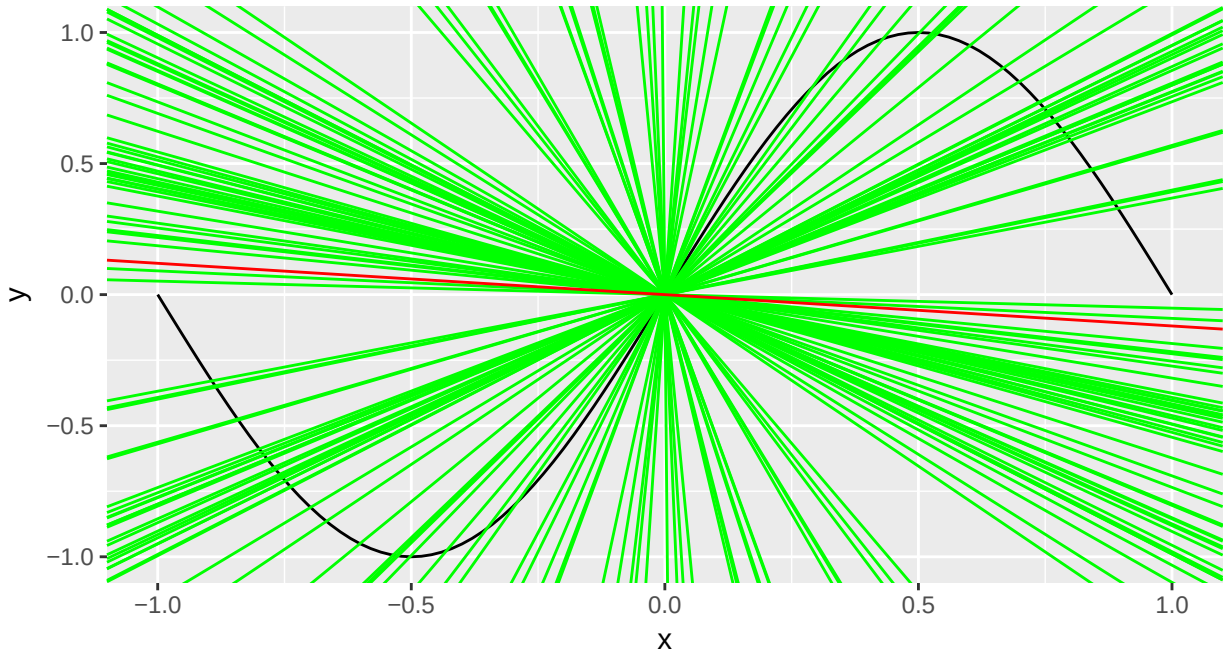
$$E_{in}(h) = \frac{1}{2}\sum_{i=1}^{2}(ax_i - y_i)^2$$

with respect to a. A simple computation shows that the best hypothesis is given by

$$g(x) = \frac{y_1x_1 + y_2x_2}{x_1 + x_2}x.$$

To compute $\overline{g}$, we will generate many training sets and evaluate a for each one of these before averaging them on these training sets to obtain $\overline{a}$ and consequently $\overline{g}(x) = \overline{a}x$.

```
plot <- ggplot(data.frame(x = seq(-1, 1, 0.001)), aes(x = x)) +
  stat_function(fun = f, geom = "line")

set.seed(1975)
a.hat <- numeric()
for (i in 1:10000) {
  rand <- runif(2, min = -1, max = 1)
  data <- data.frame(x = rand, y = f(rand))
  a <- (data$y[1] * data$x[1] + data$y[2] * data$x[2]) / (data$x[1] + data$x[2])
  a.hat <- c(a.hat, a)
  if (i %% 100 == 0)
    plot <- plot + geom_abline(intercept = 0, slope = a, col = "green")
}
a.bar <- mean(a.hat)
plot + geom_abline(intercept = 0, slope = a.bar, col = "red")
```



Now, we are able to compute the expected out-of-sample error and its bias and var components by generating a test set and averaging x on this new set.

```
x.new <- runif(100000, min = -1, max = 1)
bias.x <- (a.bar * x.new - f(x.new))^2
bias <- round(mean(bias.x), 2)
var.x <- mean((a.hat - a.bar)^2) * x.new^2
var <- round(mean(var.x), 2)
```

So, here we get a bias of 0.58 and a variance of 2494.25, which gives us an expected out-of sample error of 2494.83 which is way higher than what we got in (a).

(c) Here, we consider the learning model of all hypotheses of the form $h(x) = b$. First, to find the best hypothesis that approximates f, we use the training set $\mathcal{D} = \{(x_1, y_1), (x_2, y_2)\}$ to minimize the in-sample mean squared error

$$E_{in}(h) = \frac{1}{2}\sum_{i=1}^{2}(b - y_i)^2$$

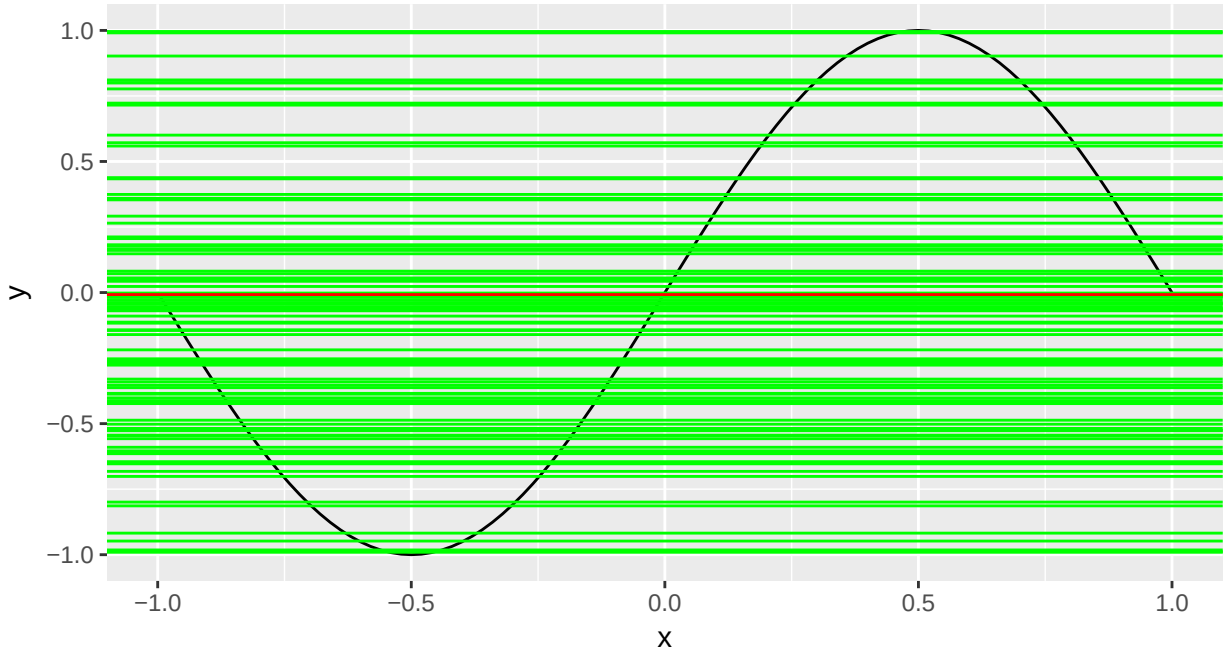
with respect to b. A simple computation shows that the best hypothesis is given by

$$g(x) = \frac{y_1 + y_2}{2},$$

which is the horizontal line at the midpointof (x_1, y_1) and (x_2, y_2). To compute $\overline{g}$, we will generate many training sets and evaluate b for each one of these before averaging them on these training sets to obtain $\overline{b}$ and consequently $\overline{g}(x) = \overline{b}$.

```
plot <- ggplot(data.frame(x = seq(-1, 1, 0.001)), aes(x = x)) +
  stat_function(fun = f, geom = "line")

set.seed(1975)
b.hat <- numeric()
for (i in 1:10000) {
  rand <- runif(2, min = -1, max = 1)
  data <- data.frame(x = rand, y = f(rand))
  b <- (data$y[1] + data$y[2]) / 2
```

```
  b.hat <- c(b.hat, b)
  if (i %% 100 == 0)
    plot <- plot + geom_abline(intercept = b, slope = 0, col = "green")
}
b.bar <- mean(b.hat)
plot + geom_abline(intercept = b.bar, slope = 0, col = "red")
```



Now, we are able to compute the expected out-of-sample error and its bias and var components by generating a test set and averaging x on this new set.

```
x.new <- runif(100000, min = -1, max = 1)
bias.x <- (b.bar - f(x.new))^2
bias <- round(mean(bias.x), 2)
var.x <- mean((b.hat - b.bar)^2)
var <- round(mean(var.x), 2)
```

So, here we get a bias of 0.5 and a variance of 0.25, which gives us an expected out-of sample error of 0.75 which is the best of the three cases we examined.

Problem 2.24

(a) We give the analytic expression for the average function $\bar{g}(x)$ below. We have

$$
\begin{aligned}
\overline{g}(x) &= \mathbb{E}_{\mathcal{D}}[g(x)] \\
&= \mathbb{E}_{\mathcal{D}}\left[\frac{y_1 - y_2}{x_1 - x_2}x + \frac{x_1 y_2 - x_2 y_1}{x_1 - x_2}\right] \\
&= \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1}\frac{x_1^2 - x_2^2}{x_1 - x_2}dx_1 dx_2 \cdot x + \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1}\frac{x_1 x_2^2 - x_2 x_1^2}{x_1 - x_2}dx_1 dx_2 \\
&= \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1}(x_1 + x_2)dx_1 dx_2 \cdot x - \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1}(x_1 x_2)dx_1 dx_2 \\
&= \frac{1}{4}\cdot 0 - \frac{1}{4}\cdot 0 = 0.
\end{aligned}
$$

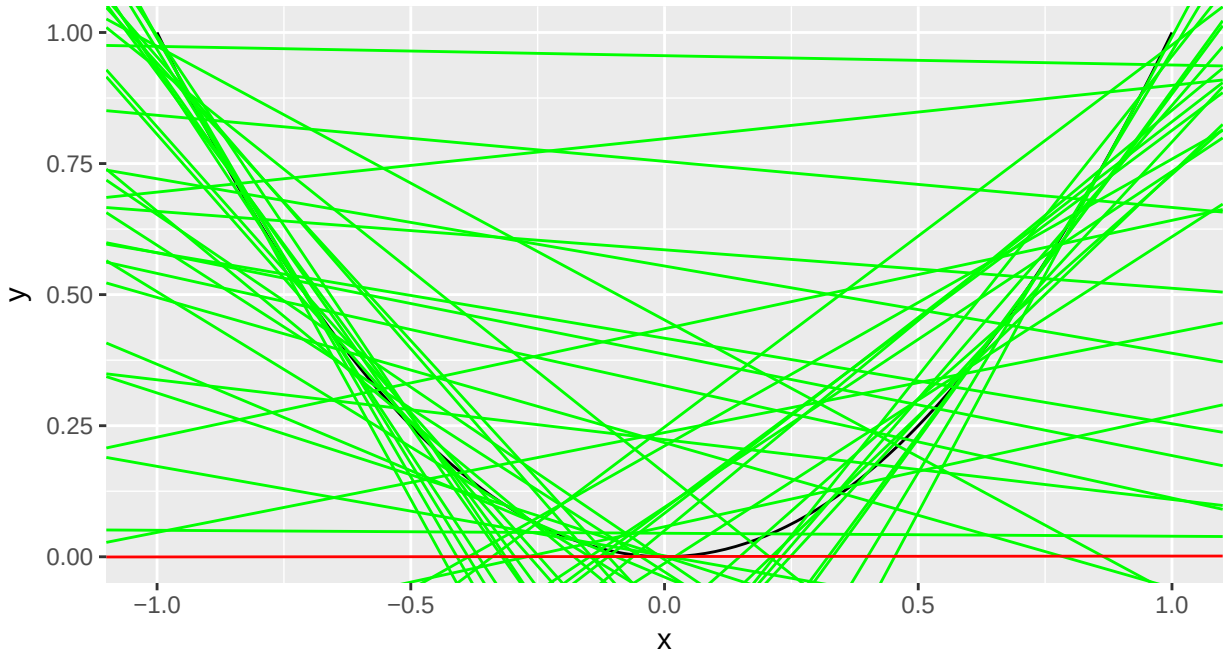
(b) To determine numerically $\overline{g}(x)$, $\mathbb{E}[E_{out}]$, the bias and variance, we design the following experiment.

```
f <- function(x)
  x^2
plot <- ggplot(data.frame(x = seq(-1, 1, 0.001)), aes(x = x)) +
  stat_function(fun = f, geom = "line")

set.seed(1975)
a.hat <- numeric()
b.hat <- numeric()
for (i in 1:50000) {
  rand <- runif(2, min = -1, max = 1)
  data <- data.frame(x = rand, y = f(rand))
  a <- ((data$y[1] - data$y[2]) / (data$x[1] - data$x[2]))
  a.hat <- c(a.hat, a)
  b <- (data$y[2] * data$x[1] - data$y[1] * data$x[2]) / (data$x[1] - data$x[2])
  b.hat <- c(b.hat, b)
  if (i %% 1000 == 0)
    plot <- plot + geom_abline(intercept = b, slope = a, col = "green")
}
a.bar <- mean(a.hat)
b.bar <- mean(b.hat)

x.new <- runif(100000, min = -1, max = 1)
Eout <- mean(x.new^4) - 2 * a.hat * mean(x.new^3) +
  (a.hat^2 - 2 * b.hat) * mean(x.new^2) +
  2 * a.hat * b.hat * mean(x.new) + b.hat^2
Exp.Eout <- round(mean(Eout), 2)
bias.x <- (a.bar * x.new - b.bar - f(x.new))^2
bias <- round(mean(bias.x), 2)
var.x <- mean((a.hat - a.bar)^2) * x.new^2 +
  2 * mean((a.hat - a.bar) * (b.hat - b.bar)) * x.new +
  mean((b.hat - b.bar)^2)
var <- round(mean(var.x), 2)
```

(c) When we run the experiment above, we get $\mathbb{E}[E_{out}] = 0.53$, a bias of 0.2, and a variance of 0.33. We may immediately see that

$$\mathbb{E}[E_{out}] = 0.53 = bias + var.$$

Below, we give a plot of $\overline{g}(x)$ and $f(x)$ together.



(d) To compute $\mathbb{E}[E_{out}]$, we will first determine E_{out}, we get

$$\begin{aligned} E_{out} = \mathbb{E}_x[(g(x) - f(x))^2] &= \mathbb{E}_x[(ax + b - x^2)^2] \\ &= \mathbb{E}_x[x^4] - 2a\mathbb{E}_x[x^3] + (a^2 - 2b)\mathbb{E}_x[x^2] + 2ab\mathbb{E}_x[x] + b^2 \\ &= \frac{1}{2}\int_{-1}^{1} x^4 dx - 2a\frac{1}{2}\int_{-1}^{1} x^3 dx + (a^2 - 2b)\frac{1}{2}\int_{-1}^{1} x^2 dx + 2ab\frac{1}{2}\int_{-1}^{1} x dx + b^2 \\ &= \frac{1}{5} + \frac{(a^2 - 2b)}{3} + b^2. \end{aligned}$$

Then, we take the expectation with respect to $\mathcal{D}$ to get the test performance and we replace a and b by $(x_1 + x_2)$ and $(-x_1 x_2)$ respectively, we get

$$\begin{aligned} \mathbb{E}_{\mathcal{D}}[E_{out}] &= \frac{1}{5} + \frac{1}{3}\mathbb{E}_{\mathcal{D}}[(x_1 + x_2)^2 + 2x_1 x_2] + \mathbb{E}_{\mathcal{D}}[x_1^2 x_2^2] \\ &= \frac{1}{5} + \frac{1}{3}\cdot\frac{1}{4}\int_{-1}^{1}\int_{-1}^{1}(x_1^2 + x_2^2 + 4x_1 x_2)dx_1 dx_2 + \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1} x_1^2 x_2^2 dx_1 dx_2 \\ &= \frac{1}{5} + \frac{1}{3}\cdot\frac{1}{4}\cdot\frac{8}{3} + \frac{1}{4}\cdot\frac{4}{9} = \frac{8}{15}. \end{aligned}$$

Next, we compute the bias, we first have

$$bias(x) = (\overline{g}(x) - f(x))^2 = f(x)^2 = x^4;$$

then we get

$$bias = \mathbb{E}_x[x^4] = \frac{1}{2}\int_{-1}^{1} x^4 dx = \frac{1}{5}.$$

Finally, we compute the variance, we first have

$$
\begin{aligned}
var(x) &= \mathbb{E}_{\mathcal{D}}[(g(x) - \overline{g}(x))^2] = \mathbb{E}_{\mathcal{D}}[a^2x^2 + 2abx + b^2] \\
&= \mathbb{E}_{\mathcal{D}}[a^2] \cdot x^2 + 2\mathbb{E}_{\mathcal{D}}[ab] \cdot x + \mathbb{E}_{\mathcal{D}}[b^2] \\
&= \mathbb{E}_{\mathcal{D}}[(x_1 + x_2)^2] \cdot x^2 - 2\mathbb{E}_{\mathcal{D}}[(x_1 + x_2)x_1x_2] \cdot x + \mathbb{E}_{\mathcal{D}}[x_1^2x_2^2] \\
&= \mathbb{E}_{\mathcal{D}}[x_1^2 + 2x_1x_2 + x_2^2] \cdot x^2 - 2\mathbb{E}_{\mathcal{D}}[x_1^2x_2 + x_1x_2^2] \cdot x + \mathbb{E}_{\mathcal{D}}[x_1^2x_2^2] \\
&= \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1}(x_1^2 + 2x_1x_2 + x_2^2)dx_1dx_2 \cdot x^2 - \frac{2}{4}\int_{-1}^{1}\int_{-1}^{1}(x_1^2x_2 + x_1x_2^2)dx_1dx_2 \cdot x + \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1}x_1^2x_2^2dx_1dx_2 \\
&= \frac{1}{4}(\frac{4}{3} + 0 + \frac{4}{3}) \cdot x^2 - 0 \cdot x + \frac{1}{4} \cdot \frac{4}{9} = \frac{2}{3} \cdot x^2 + \frac{1}{9};
\end{aligned}
$$

then we get

$$
var = \mathbb{E}_x\left[\frac{2}{3}x^2 + \frac{1}{9}\right] = \frac{2}{3} \cdot \frac{1}{2}\int_{-1}^{1}x^2dx + \frac{1}{9} = \frac{1}{3}.
$$

We may see that these analytical results are pretty close to the numerically obtained ones.