Homework 7 - Solution

1. **Proving that** $\mathrm{VC}(\mathcal{F}) = 4$ **rigorously was not needed here.** There exists a set of 4 points that can be shattered by the set of rectangles (see Figure 1) so $\mathrm{VC}(\mathcal{F}) \geq 4$. Let us show that any set of 5 points cannot be shattered by $\mathcal{F}$, implying that we have $\mathrm{VC}(\mathcal{F}) = 4$. Let $x_1, \ldots, x_5 \in \mathbb{R}^2$, and assume without loss of generality that $x_1^{(1)} \leq x_2^{(1)} \leq \cdots \leq x_5^{(1)}$. We also assume for the sake of simplicity that all the horizontal coordinates $x_i^{(1)}$ are distinct, as well as all the vertical coordinates $x_i^{(2)}$. Let $\phi(i)$ be the index of the ith smallest vertical coordinates among the x_is, so that
$$x_{\phi(1)}^{(2)} < x_{\phi(2)}^{(2)} < \cdots < x_{\phi(5)}^{(2)}.$$
The function ϕ defines a permutation of $\{1, \ldots, 5\}$. One can check that there always exist three indices $i_1 < i_2 < i_3$ with $\phi(i_1) < \phi(i_2) < \phi(i_3)$ or $\phi(i_3) < \phi(i_2) < \phi(i_1)$. It is not possible to select only x_{i_1} and x_{i_3} with a rectangle (indeed such a rectangle will also contain x_{i_2}). Therefore, $\mathcal{F}$ does not shatter the set of inputs.

2. (a) Any function $h \in \mathcal{H}$ is of the form $h(x) = \max_i f_i(x)$ for some function $f_i \in \mathcal{F}_i$. Therefore, we can define a surjection from $\mathcal{F}_1 \times \cdots \times \mathcal{F}_k$ to $\mathcal{H}$. Given inputs $x_1, \ldots, x_n \in \mathcal{X}$, this yields a surjection from $\mathcal{C}_{\mathcal{F}_1}(x_1, \ldots, x_n) \times \cdots \times \mathcal{C}_{\mathcal{F}_k}(x_1, \ldots, x_n)$ to $\mathcal{C}_{\mathcal{H}}(x_1, \ldots, x_n)$. This implies that
$$\mathcal{N}_{\mathcal{H}}(x_1, \ldots, x_n) \leq \prod_{i=1}^{k} \mathcal{N}_{\mathcal{F}_i}(x_1, \ldots, x_n)$$
and we conclude by taking applying the log function to both sides of this inequality.

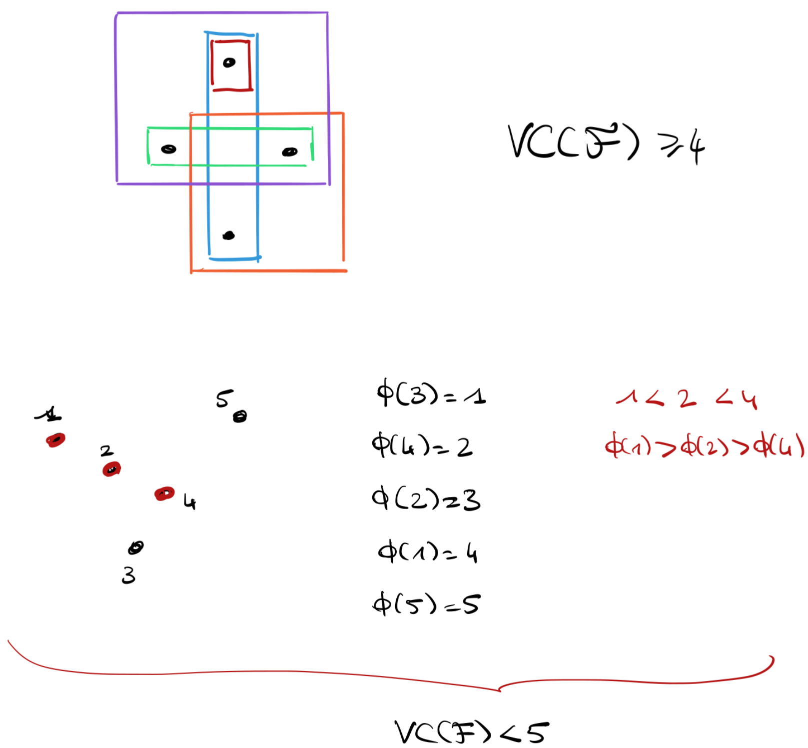

Figure 1: Top: Example of a set of four points that can be shattered by axis-aligned rectangles. Bottom: A set of five points cannot be shattered. The middle red point cannot be separated from the two other red points.

(b) Let $n = CDk\log(k)$ for some constant C to be fixed. As long as $C > 1/\log(2)$, we have $n > 2D$. Therefore, we can apply Sauer's lemma to each of the sets $\mathcal{F}_i$ (and use that the function $x \mapsto \log(en/x)$ is increasing on $[0, n]$):

$$\begin{aligned}\log(\mathcal{N}_{\mathcal{F}_i}(x_1, \dots, x_n)) &\leq \mathrm{VC}(\mathcal{F}_i)\log\left(\frac{en}{\mathrm{VC}(\mathcal{F}_i)}\right)\\ &\leq D\log\left(\frac{en}{D}\right) \leq D\log(Cek\log(k)).\end{aligned}$$

Using the previous question, we obtain that

$$\log(\mathcal{N}_{\mathcal{H}}(x_1, \dots, x_n)) \leq kD\log(Cek\log(k)).$$

By the definition of the VC dimension, if $\log(\mathcal{N}_{\mathcal{H}}(x_1, \dots, x_n)) < n\log(2)$ for every inputs $x_1, \dots, x_n$, then $\mathrm{VC}(\mathcal{H}) < n$. However, we have

$$n\log(2) = C\log(2)Dk\log(k).$$

To conclude, we choose C such that

$$C\log(2)\log(k) > \log(Cek\log(k))$$

for every $k \geq 2$. One can check that $C = 7$ is enough for instance. Therefore,

$$\mathrm{VC}(\mathcal{H}) < 7Dk\log(k).$$

3. (a) Define

$$f_0(x) = \begin{cases} 1 & \text{if } x^{(2)} \leq g_0(x^{(1)}) \\ -1 & \text{otherwise.} \end{cases} \tag{1}$$

The Bayes risk $\mathcal{R}_P(f_0)$ is equal to $P(f_0(\mathbf{x}) \neq \mathbf{y}) = 0$ by definition of $\mathbf{y}$. As we have $\mathcal{R}_P(f) \geq 0$ for any function f, this implies both that $f_P^\star = f_0$ and that $\mathcal{R}_P(f_P^\star) = 0$.

(b) The Bayes predictor $f_P^\star$ belongs to $\mathcal{F}$. Also, for every observation $(\mathbf{x_i}, \mathbf{y_i})$, we have $f_P^\star(\mathbf{x_i}) = \mathbf{y_i}$ by definition. Therefore, $\mathcal{R}_n(f_P^\star) = 0$ and $f_P^\star$ is a minimizer of $\mathcal{R}_n$. However, there are many functions $f \in \mathcal{F}$ with $\mathcal{R}_n(f) = 0$, so that there is no uniqueness of the empirical risk minimizer! The approximation error $\inf_{f\in\mathcal{F}} \mathcal{R}_P(f) -$

$\mathcal{R}_P(f_P^\star)$ is equal to $0 - 0 = 0$. The main issue with this predictor is that, in practice, we will select one of the many functions f that satisfy $\mathcal{R}_n(f) = 0$ as our predictor, and nothing tells us that this predictor is close from $f_P^\star$. The estimation error will likely be very large for this set $\mathcal{F}$.

(c) Let $x \in [l/L, (l+1)/L)$. We have

$$\begin{aligned} g_0(x) &= g_0(x_l) + g_0'(x_l)(x - x_l) + \cdots + g_0^{(k-1)}(x_l)\frac{(x - x_l)^{k-1}}{(k-1)!} \\ &\quad + \int_{x_l}^{x} \frac{g_0^{(k)}(t)}{(k-1)!}(x-t)^{k-1}\mathrm{d}t \\ &= \tilde{g}_{0,l}(x) + \int_{x_l}^{x} \frac{g_0^{(k)}(t)}{(k-1)!}(x-t)^{k-1}\mathrm{d}t. \end{aligned}$$

The function $\tilde{g}_{0,l}$ is a polynomial function of degree $k-1$. The remainder integral term is bounded by $\frac{R}{k!}|x - x_l|^k \leq \frac{R}{k!(2L)^k}$. We can define a function $\tilde{g}_0$ in $\mathcal{G}_{l,k}$ by letting $\tilde{g}_0(x) = \tilde{g}_{0,l}(x)$ if $x \in [l/L, (l+1)/L)$. Consider the associated classifier $\tilde{f}_0$. The approximation error is bounded by

$$\mathcal{R}_P(\tilde{f}_0) - \mathcal{R}_P(f_P^\star) = \mathcal{R}_P(\tilde{f}_0).$$

Also,

$$\begin{aligned} \mathcal{R}_P(\tilde{f}_0) &= P(\tilde{f}_0(\mathbf{x}) \neq \mathbf{y}) \\ &= P(\tilde{g}_0(\mathbf{x}^{(\mathbf{1})}) < \mathbf{x}^{(\mathbf{2})} \text{ and } g_0(\mathbf{x}^{(\mathbf{1})}) \geq \mathbf{x}^{(\mathbf{2})}) \\ &\quad + P(\tilde{g}_0(\mathbf{x}^{(\mathbf{1})}) \geq \mathbf{x}^{(\mathbf{2})} \text{ and } g_0(\mathbf{x}^{(\mathbf{1})}) < \mathbf{x}^{(\mathbf{2})}). \end{aligned}$$

This sum is equal to the probability that $\mathbf{x}^{(\mathbf{2})}$ is between $\tilde{g}_0(\mathbf{x}^{(\mathbf{1})})$ and $g_0(\mathbf{x}^{(\mathbf{1})})$. As $\mathbf{x}$ is uniform in $[0,1]^2$, we have

$$\begin{aligned} \mathcal{R}_P(\tilde{f}_0) &= \mathbb{E}[|\tilde{g}_0(\mathbf{x}^{(\mathbf{1})}) - g_0(\mathbf{x}^{(\mathbf{1})})|] \\ &\leq \frac{R}{k!(2L)^k}. \end{aligned}$$

This is our final bound on the approximation error.

(d) **No proof required here. However, here is a geometrical proof for those interested.**

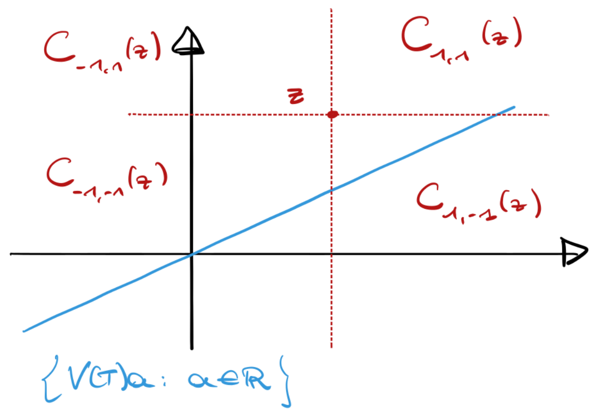

Figure 2: The quadrant $C_{-1,1}(z)$ does not intersect the image of $V(T)$. This implies that we cannot obtain the classification $-1, +1$ with the inputs corresponding to T and z.

- The VC dimension of $\mathcal{F}_{1,1}$ is 1: with a set of two inputs x_1, x_2, with for instance $x_1^{(2)} \leq x_2^{(2)}$, then we cannot assign x_1 to -1 and x_2 to $+1$ with a classifier in $\mathcal{F}_{1,1}$.
- The VC dimension of $\mathcal{F}_{1,k}$ is k. This can maybe best be seen geometrically. Let $T = (t_1, \ldots, t_l)$ be a set of l numbers between 0 and 1. We consider the matrix $V[T]$ of size $l \times k$, with $V[T]_{i,j} = t_i^{j-1}$. If $a = (a_0, \ldots, a_{k-1}) \in \mathbb{R}^k$ then $V[T]a \in \mathbb{R}^l$ is equal to $(g(t_1), \ldots, g(t_l))$, where $g(t) = \sum_{j=1}^{k} a_j t^{j-1} \in \mathcal{G}_{1,k}$.
  Let us understand what it means that $\mathcal{G}_{1,k}$ shatters a set of l inputs $(x_1, \ldots, x_l)$. Let $t_i = x_i^{(1)}$ and $z_i = x_i^{(2)}$. Consider the vector $z = (z_1, \ldots, z_l) \in \mathbb{R}^l$. To obtain the classification $y = (y_1, \ldots, y_l)$ associated with the classifier g (corresponding to a vector $a \in \mathbb{R}^k$), we have to consider the relative position of the vectors $u = V[T]a$ and z. The sign y_i will be $+1$ if $z_i \leq u_i$, and -1 otherwise. The vectors a leading to a classification y are exactly the vectors such that $u = V[T]a$ belongs to

  $$C_y(z) = \{u : \ u_i \geq z_i \text{ if } y_i = +1 \text{ and } u_i < z_i \text{ otherwise}\}.$$

  Each of the set $C_y(z)$ is a "quadrant" centered at z. The set

$\mathcal{G}_{1,k}$ shatters $(x_1, \ldots, x_l)$ exactly if, for all 2^l possible configurations of signs $y = (y_1, \ldots, y_l)$, there exists $a \in \mathbb{R}^l$ with $V[T]a \in C_y(z)$. See also Figure 2.

If $l = k$, then we can find $x_1, \ldots, x_l$ such that the matrix $V[T]$ is of rank k. Then, the image set $\{V[T]a,\ a \in \mathbb{R}^l\}$ is equal to $\mathbb{R}^l$. In particular, we can find a vector a such that $V[T]a \in C_z(y)$ for any choice of signs $y = (y_1, \ldots, y_l)$, implying that $x_1, \ldots, x_l$ is shattered. Therefore, $\mathrm{VC}(\mathcal{G}_{1,k}) \geq k$.

If $l = k + 1$, then, for any inputs $x_1, \ldots, x_{l+1}$, the rank of the matrix $V[T]$ is at most k. Therefore, the image set $\{V[T]a,\ a \in \mathbb{R}^l\}$ is a subspace of dimension at most k of $\mathbb{R}^{k+1}$. In particular, the image set does not intersect all quadrants $C_y(z)$.

- The restrictions of a function $g \in \mathcal{G}_{L,0}$ to each interval $[l/L, (l+1)/L)$ can be chosen independently. This implies that

$$\mathrm{VC}(\mathcal{F}_{L,1}) = L\mathrm{VC}(\mathcal{F}_{1,1}) = L.$$

- Likewise, $\mathrm{VC}(\mathcal{F}_{L,k}) = Lk$.

(e) According to Theorem 3.7 in the lecture notes, the expected estimation error is bounded by

$$2\sqrt{\frac{2Lk}{n}\log\left(\frac{en}{Lk}\right)}.$$

(f) The expected excess of risk $\mathbb{E}[\mathcal{R}_P(\hat{f}_{\mathcal{F}_{L,k}}) - \mathcal{R}_P(f_P^\star)]$ is bounded by the sum of the approximation error and of the expected estimation error, which is bounded by

$$\frac{R}{k!(2L)^k} + 2\sqrt{\frac{2Lk}{n}\log\left(\frac{en}{Lk}\right)}.$$

Forgetting about the log factors, this is the sum of a quantity decreasing in L (of order L^{-k}) and of a quantity increasing in L (of order $\sqrt{L/n}$). The minimum of $L^{-k} + \sqrt{L/n}$ is attained for $L = n^{1/(2k+1)}$, and yields an expected excess of risk of order $n^{-k/(2k+1)}$.