A. Rademacher complexity

1. Consider the class of functions $\mathcal{H}$ mapping from $\mathbb{R}$ to $\{+1, -1\}$ such that

$$h(x) = \begin{cases} +1 \text{ for } x \in [a, b], \\ -1 \text{ otherwise }, \end{cases}$$

for some $a, b \in \mathbb{R}$. Give an upper bound on the growth function $\Pi_{\mathcal{H}}(m)$ and use it to derive an upper bound on the $\mathfrak{R}_m(\mathcal{H})$.

Solution:
By definition, $\mathcal{H}$ is the hypothesis class of intervals in the real line and therefore

$$\text{VCdim}(\mathcal{H}) = d = 2$$

Plugging in this value into *Sauer's Lemma* (corollary 3.3 in the textbook) we get:

$$\Pi_{\mathcal{H}}(m) \leq \left(\frac{em}{2}\right)^2$$

Finally we use *Massart's Lemma* (corollary 3.1 in the textbook) to upper bound $\mathfrak{R}_m(\mathcal{H})$:

$$\begin{aligned} \mathfrak{R}_m(\mathcal{H}) &\leq \sqrt{\frac{2log\Pi_{\mathcal{H}}(m)}{m}} \\ &\leq \sqrt{\frac{2log\left[\left(\frac{em}{2}\right)^2\right]}{m}} \\ &\leq 2\sqrt{\frac{log\left(\frac{em}{2}\right)}{m}} \end{aligned}$$

2. Prove that for any hypotheses class $\mathcal{H}$ and any function $h_0 \colon \mathcal{X} \mapsto \mathbb{R}$, $\mathfrak{R}_m(\mathcal{H}) = \mathfrak{R}_m(\mathcal{H} + h_0)$.

Solution:

$$\begin{aligned} \mathfrak{R}_m(\mathcal{H} + h_0) &= \underset{S \sim D^m}{E}\left[\widehat{\mathfrak{R}}_S(\mathcal{H} + h_0)\right] \\ &= \underset{S}{E}\left[\underset{\sigma}{E}\left[\sup_{h \in \mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i[h(z_i) + h_0(z_i)]\right)\right]\right] \\ &= \underset{S,\sigma}{E}\left[\sup_{h \in \mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(z_i)\right)\right] + \underset{S,\sigma}{E}\left[\frac{1}{m}\sum_{i=1}^{m}\sigma_i h_0(z_i)\right] \\ &= \mathfrak{R}_m(\mathcal{H}) + \underset{S}{E}\left[\frac{1}{m}\sum_{i=1}^{m}\underset{\sigma}{E}[\sigma_i]h_0(z_i)\right] \\ &= \mathfrak{R}_m(\mathcal{H}) \quad (Q.E.D.) \end{aligned}$$

- The first step is straightforward from the definition of Rademacher complexity for $\mathcal{H}' = \mathcal{H} + h_0$.
- The second step follows from the linearity of expectation and linearity of supremum (a proof for the linearity of supremum can be found here). Also, the second summation is constant with respect to h so we can get rid of the supremum.
- The third step uses the definition of Rademacher complexity for $\mathcal{H}$ and linearity of expectation.
- Finally, the last step follows from the fact that $E_\sigma[\sigma_i] = 0$ since the σ_i are Rademacher variables.

3. Prove that if for two hypotheses classes $\mathcal{H}$ and $\mathcal{F}$ the inclusion $\mathcal{H} \subseteq \mathcal{F}$ holds, then the following inequality holds for any finite sample S: $\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq \widehat{\mathfrak{R}}_S(\mathcal{F})$.

Solution:
Let us define the following sets:

$$S_{\mathcal{H}} = \left\{ \sum_{i=1}^{m} \sigma_i h(z_i) : h \in \mathcal{H} \right\} \quad S_{\mathcal{F}} = \left\{ \sum_{i=1}^{m} \sigma_i f(z_i) : f \in \mathcal{F} \right\}$$

Given the definition of the Rademacher complexity, proving that $\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq \widehat{\mathfrak{R}}_S(\mathcal{F})$ is equivalent to proving $\sup(S_{\mathcal{H}}) \leq \sup(S_{\mathcal{F}})$. Since we are considering a finite sample of size m, both suprema must exist. Furthermore, since we know $\mathcal{H} \subseteq \mathcal{F}$, it must follow that $S_{\mathcal{H}} \subseteq S_{\mathcal{F}}$.
Now we suppose that $\sup(S_{\mathcal{H}}) > \sup(S_{\mathcal{F}})$. Under this assumption, it must exist an element $e \in S_{\mathcal{H}}$ such that $e > \sup(S_{\mathcal{F}})$. But if $e \in S_{\mathcal{H}}$ and $S_{\mathcal{H}} \subseteq S_{\mathcal{F}}$, it must follow that $e \in S_{\mathcal{F}}$ which contradicts the definition of $\sup(S_{\mathcal{F}})$. Therefore, it must be that $\sup(S_{\mathcal{H}}) \leq \sup(S_{\mathcal{F}})$ which implies $\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq \widehat{\mathfrak{R}}_S(\mathcal{F})$ ***(Q.E.D.)***
Intuitively, if we plug $\mathcal{H}$ into the definition of Rademacher complexity, the supremum is taken over all the functions in $\mathcal{H}$. On the other hand if we consider the Rademacher complexity of $\mathcal{F}$, the supremum will increase (or at least remain the same) because it is taken over a larger domain.

4. Let $\mathcal{H}_1$ be a family of functions mapping $\mathcal{X}$ to $\{0, 1\}$ and let $\mathcal{H}_2$ be a family of functions mapping $\mathcal{X}$ to $\{-1, +1\}$. Let $\mathcal{H} = \{h_1 h_2 \colon h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$. Show that the empirical Rademacher complexity of $\mathcal{H}$ for any sample S of size m can be bounded as follows:

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq \widehat{\mathfrak{R}}_S(\mathcal{H}_1) + \widehat{\mathfrak{R}}_S(\mathcal{H}_2).$$

Solution:
First we rewrite $h_1 h_2$ as follows:

$$h_1 h_2 = \frac{1}{2} \left[|h_1 + h_2| - |h_1 - h_2| \right]$$

And with this new definition for $h \in \mathcal{H}$ we write:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{H}) &= \underset{\sigma}{E} \left[\sup_{h \in \mathcal{H}} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(\frac{1}{2} \left[|h_1 + h_2| - |h_1 - h_2| \right] \right) \right) \right] \\
&= \underset{\sigma}{E} \left[\sup_{h \in \mathcal{H}} \left(\frac{1}{2m} \sum_{i=1}^{m} \sigma_i \left| h_1 + h_2 \right| \right) + \sup_{h \in \mathcal{H}} \left(-\frac{1}{2m} \sum_{i=1}^{m} \sigma_i \left| h_1 - h_2 \right| \right) \right] \\
&= \underset{\sigma}{E} \left[\sup_{h \in \mathcal{H}} \left(\frac{1}{2m} \sum_{i=1}^{m} \sigma_i \left| h_1 + h_2 \right| \right) \right] + \underset{\sigma}{E} \left[\sup_{h \in \mathcal{H}} \left(-\frac{1}{2m} \sum_{i=1}^{m} \sigma_i \left| h_1 - h_2 \right| \right) \right] \\
&\leq \underset{\sigma}{E} \left[\sup_{h \in \mathcal{H}} \left(\frac{1}{2m} \sum_{i=1}^{m} \sigma_i \left(h_1 + h_2 \right) \right) \right] + \underset{\sigma}{E} \left[\sup_{h \in \mathcal{H}} \left(-\frac{1}{2m} \sum_{i=1}^{m} \sigma_i \left(h_1 - h_2 \right) \right) \right] \\
&\leq \frac{1}{2} \left[\widehat{\mathfrak{R}}_S(\mathcal{H}_1 + \mathcal{H}_2) \right] + \frac{1}{2} \left[\widehat{\mathfrak{R}}_S(\mathcal{H}_1 - \mathcal{H}_2) \right] \\
&\leq \widehat{\mathfrak{R}}_S(\mathcal{H}_1) + \widehat{\mathfrak{R}}_S(\mathcal{H}_2) \quad (Q.E.D.)
\end{aligned}$$

- The first three steps are straightforward by the linearity of summation, supremum and expectation.
- In the fourth step we introduce the inequality by bounding the absolute value function (which is 1-Lipschitz) using *Talagrand's inequality* (lemma 4.2 in the textbook).
- Finally, we apply the Rademacher complexity definition and two basic properties to complete the proof: linearity $\mathfrak{R}(\mathcal{H}_1 + \mathcal{H}_2) = \mathfrak{R}(\mathcal{H}_1) + \mathfrak{R}(\mathcal{H}_2)$, which must hold since $\mathfrak{R}$ is defined in terms of linear operators supremum and summation, and multiplication by a scalar $\mathfrak{R}(\alpha\mathcal{H}) = |\alpha| \mathfrak{R}(\mathcal{H})$, which introduces the absolute value of α because the Rademacher variables σ_i and $-\sigma_i$ follow the same distribution.

B. VC-dimension

1. What is the VC-dimension of axis-aligned squares in $\mathbb{R}^2$?
   Solution:
   If we consider the set of 3 points $S = \{P_1(0,0), P_2(2d,0), P_3(d,d)\}$, it can be fully shattered by axis-aligned squares:
   - To label positively P_1 and P_2, we use the square defined by $\{P_1, P_2, (0,2d), (2d,2d)\}$.
   - To label positively P_1 and P_3, we use the square defined by $\{P_1, P_3, (0,d), (d,0)\}$.
   - To label positively P_2 and P_3, we use the square defined by $\{P_2, P_3, (2d,d), (d,0)\}$.
   - The remaining 5 labelings are straightforward.

   So the VC-dimension must be at least 3. To prove that no set of 4 points can be fully shattered we do the following:
   - Let P_{maxY} be the point with the bigger y-coordinate and P_{minY} the point with the smaller y-coordinate. Similarly we define P_{maxX} and P_{minX} and assume that there are no ties, i.e. the 4 points can be defined in a unique way.
   - Let d_y be the vertical distance between P_{maxY} and P_{minY}, and d_x the horizontal distance between P_{maxX} and P_{minX}.
   - If $d_x \geq d_y$, the labeling with P_{maxX} and P_{minX} positive is not possible, since at least one of P_{maxY} and P_{minY} would lie within the square.
   - Similarly, if $d_y \geq d_x$ the labeling with P_{maxY} and P_{minY} positive is not possible.

   The special cases for which the 4 points are not uniquely defined can be approached using the same analysis:
   - If any two points have the same value for the x or y coordinate and it is the maximum/minimum, we can arbitrarily choose any of the two as the maximum/minimum.
   - The analysis doesn't rely on the fact that $P_{maxY} \neq P_{maxX}$ or any other such relationship, so a single point can in fact be maximum/minimum in both axes.

   Since no set of 4 points can be fully shattered, the VC-dimension of axis-aligned squares in the plane must be 3.
2. What is the VC-dimension of intersections of 2 axis-aligned squares in $\mathbb{R}^2$?
   Solution:
   The intersection of two axis-aligned squares is an axis-aligned rectangle, and therefore the VC-dimension is 4 (The proof for the VC-dimension of axis-aligned rectangles was covered in class).
3. (Bonus) Let C be a concept class whose VC-dimension is 3. Show that the VC-dimension of intersections of k concepts from C is upper bounded by $6k\log_2(3k)$. (***hint***: use Sauer's lemma.)
   Solution:
   To be able to apply Sauer's lemma, we will first prove that the following inequality holds:
   $$\Pi_C(m) \leq \Pi_{C_1}(m)\Pi_{C_2}(m)$$
   Where $C = C_1 \cap C_2$. First we take any set X of m points and define k_1 and k_2 as the number of distinct subsets of X that are positively labeled by the concepts in C_1 and C_2, respectively. By definition of the growth function we have that $k_1 \leq \Pi_{C_1}(X) \leq \Pi_{C_1}(m)$ and $k_2 \leq \Pi_{C_2}(X) \leq \Pi_{C_2}(m)$.
   Now we notice that the subsets of X that are positively labeled by the concepts in C are formed by intersections of the subsets of X positively labeled by the concepts in C_1 and the subsets of X positively labeled by the concepts in C_2. Therefore, the number of distinct positively labeled subsets of X by the concepts in C satisfies $\Pi_C(X) \leq k_1 k_2 \leq \Pi_{C_1}(m)\Pi_{C_2}(m)$.
   Since this holds for all X of size m, we conclude that $\Pi_C(m) \leq \Pi_{C_1}(m)\Pi_{C_2}(m)$ ***(Q.E.D)***.
   Using this result, we define C_k to be the concept class formed by all intersections of k concepts from C and we write
   $$\Pi_{C_k}(m) \leq (\Pi_C(m))^k$$
   Which using Sauer's lemma and corollary 3.3 from the textbook (with $d = 3$) implies
   $$\Pi_{C_k}(m) \leq \left(\frac{em}{3}\right)^{3k}$$

Now we notice that if $\left(\frac{em}{3}\right)^{3k} < 2^m$, it must follow that the VC-dimension of C_k is less than m. So we just need to show that this inequality holds for $m = 6k\log_2(3k)$:

$$\begin{aligned}
\left(\frac{6ek\log_2(3k)}{3}\right)^{3k} &< 2^{6k\log_2(3k)} \\
\left(\frac{6ek\log_2(3k)}{3}\right)^{3k} &< \left(2^{2\log_2(3k)}\right)^{3k} \\
\frac{6ek\log_2(3k)}{3} &< (3k)^2 \\
\log_2(3k) &< \frac{9k}{2e}
\end{aligned}$$

Which is clearly true for all $k > 0$ and completes the proof.

C. Support Vector Machines

1. Download and install the `libsvm` software library.
2. Consider the shuffled version of the `svmguide1` dataset. Use the `libsvm` scaling tool to scale the features of all the data. Use the first 2316 examples for training and the last 773 for testing. The scaling parameters should be computed only on the training data and then applied to the test data.
3. Consider the binary classification task in `svmguide1`, using the 4 features. Randomly split the training data into ten equal-sized disjoint sets. For each value of the polynomial degree, $d = 1, 2, 3, 4$, plot the average cross-validation error plus or minus one standard deviation as a function of C varying C in powers of 2, starting from a small value $C = 2^{-k}$ to $C = 2^k$, for some values of k.
   Solution:
   Figure 1 shows the desired plot.
4. Let (C^*, d^*) be the best pair found previously. Fix C to be C^*. Plot the following results as a function of d:
   (a) The average ten-fold cross-validation error, and the test error for the hypotheses obtained by running SVMs on the whole training set.
   (b) The average number of support vectors, and the average number of support vectors that lie on the marginal hyperplanes.

   Solution:
   From the previous exercise we fix $C^* = 2^{13} = 8\,192$, which yields the best average cross-validation error $e = 0.0302245$ when $d = 3$. Using this value of C^*, Figures 2 and 3 show the desired plots. The results obtained with `libsvm` include only the total number of support vectors SV and the number of boundary support vectors BSV (associated to outliers). To calculate the number of suport vectors on the marginal hyperplanes we simply do $MSV = SV - BSV$.
5. SVMs are "sparse" in the sense that the number of support vectors is usually small compared to total number of observations. Suppose we explicitly maximize sparsity by penalizing the L_2 norm of the vector α that defines the weight vector w:

$$\begin{aligned} \min_{\alpha,b,\xi} \quad & \frac{1}{2} \parallel \alpha \parallel^2 + C \left(\sum_{i=1}^{m} \xi_i \right) \\ \text{subject to} \quad & y_i \left(\left(\sum_{j=1}^{m} \alpha_j y_j x_j \right) \cdot x_i + b \right) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \alpha_i \geq 0, i \in [1, m]. \end{aligned}$$

   Show that the problem coincides with an instance of the primal optimization problem of SVMs, modulo the non-negativity constraint on α. You should indicate exactly how to view it as such.
   Solution:
   If we introduce the auxiliary variables

$$\boldsymbol{z_i} = (x_i x_1 y_1, x_i x_2 y_2, \ldots, x_i x_m y_m)$$

   The inner product of $\boldsymbol{\alpha}$ and $\boldsymbol{z_i}$ is given by

$$\boldsymbol{\alpha} \cdot \boldsymbol{z_i} = \sum_{j=1}^{m} \alpha_j z_j = \sum_{j=1}^{m} \alpha_j x_i x_j y_j$$

   Which is exactly the expression that we have in the conditions of our original problem so we can write

$$\begin{aligned} \min_{\alpha,b,\xi} \quad & \frac{1}{2} \parallel \alpha \parallel^2 + C \sum_{i=1}^{m} \xi_i \\ \text{subject to} \quad & y_i \left(\boldsymbol{\alpha} \cdot \boldsymbol{z_i} + b \right) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \alpha_i \geq 0, i \in [1, m]. \end{aligned}$$

   Aside from the non-negativity constraints on the α_i, this new formulation of the problem coincides with the standard formulation of a primal SVM optimization problem on samples $\boldsymbol{z_i}$ *(Q.E.D.)*

Figure 1: Average cross-validation error as a function of the trade-off parameter C for different values of the degree d of the polynomial kernel.

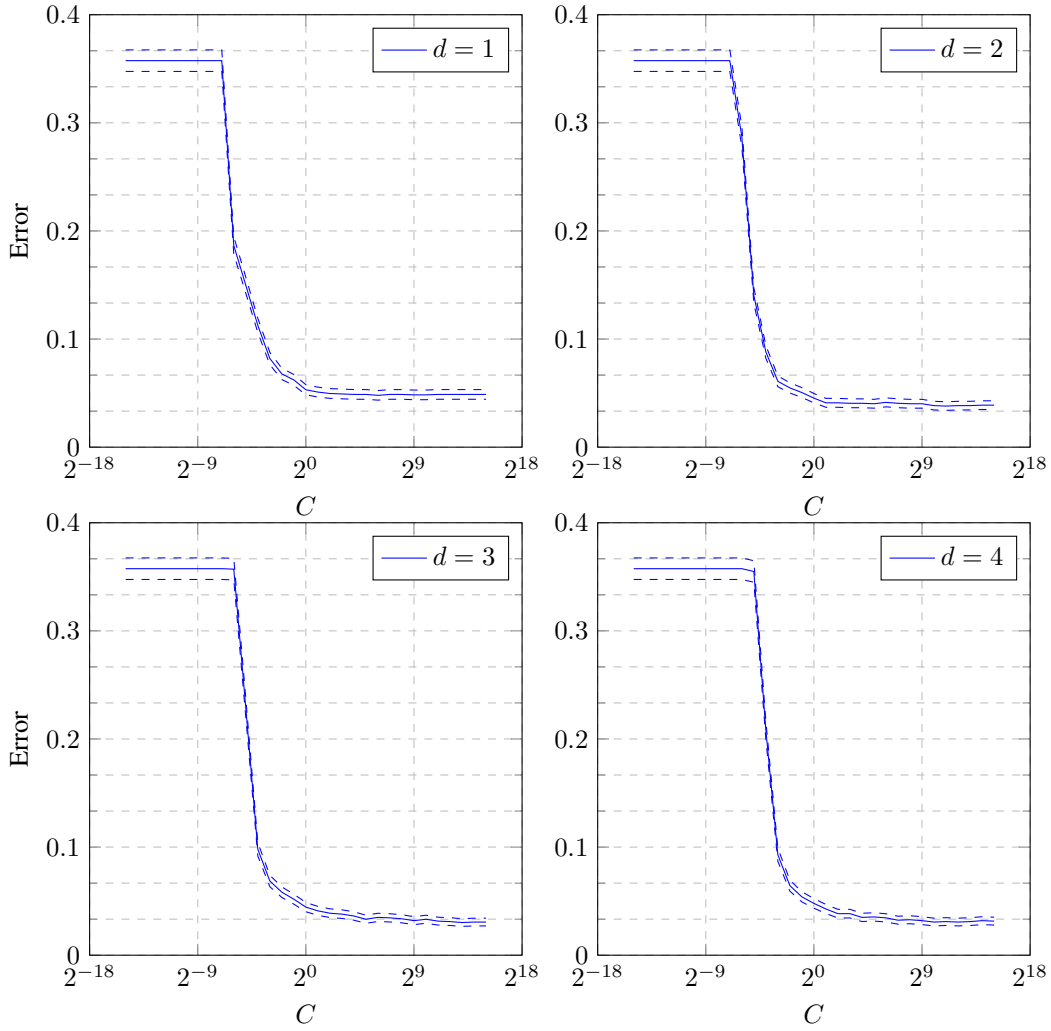

Figure 2: Training error (Mean cross-validation error) and Test error as a function of the degree d of the polynomial kernel.

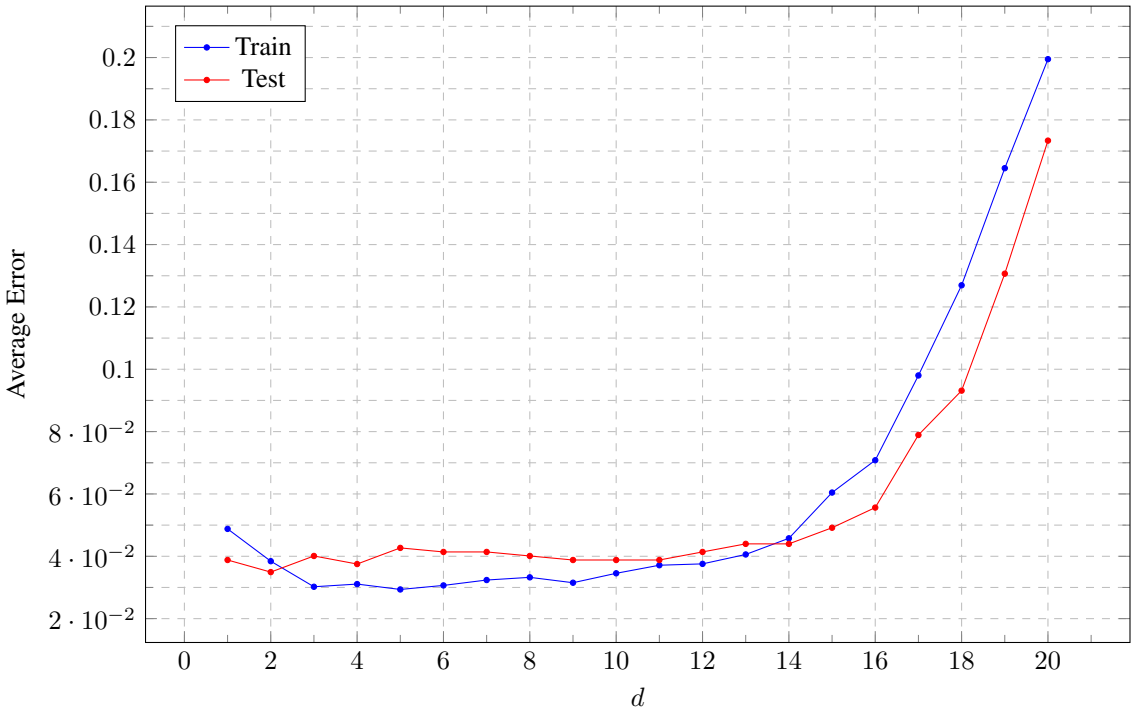

Figure 3: Support Vectors (SV) and Support Vectors on Marginal Hyperplanes (MSV) as a function of the degree d of the polynomial kernel.