A. Structural Risk Minimization

As discussed in class, the Structural Risk Minimization (SRM) technique is based on a hypothesis set $\mathcal{H}$ defined as a countable union of hypothesis sets $\mathcal{H}_n$ with finite VC-dimension or favorable Rademacher complexity. In this problem, we study several questions related to such countable union hypothesis sets.

1. Let $\mathcal{H} = \bigcup_{n=1}^{+\infty} \{h_n\}$ be a countable hypothesis set and assume that the target labeling function is in $\mathcal{H}$. In the standard statistical learning scenario, the learner receives an i.i.d. sample that he uses to train an algorithm and return a predictor. Here, suppose instead that the learner can request more labeled samples drawn i.i.d., as needed. Consider the following algorithm: starting from $t = 1$, at each round $t$, sample $m_t = \frac{1}{\epsilon} \log \frac{1}{\delta t}$ labeled points; if $h_t$ is consistent with $m_t$, return $h_t$ and stop.

   (a) Prove that the algorithm terminates.

   **Solution:** Since the Bayes classifier $f^*$ is in $\mathcal{H}$, there exists $t$ such that $f^* = h_t$, thus the algorithm terminates at most after $t$ rounds.

   (b) Fix $\epsilon, \delta > 0$ and choose $\delta_t = \frac{\delta}{2t^2}$. Show that with probability $1 - \delta$, the algorithm returns a hypothesis with error at most $\epsilon$. Suppose we use the samples obtained from previous rounds to test consistency, then, what is the maximum number of samples needed by the algorithm?

   **Solution:** The probability that the algorithm stops at round $t$ while $h_t$ has error $\epsilon$ is $\mathbb{P}[h_t \text{ consistent}\mid R(h_t) \geq \epsilon] \leq (1 - \epsilon)^{m_t} \leq e^{-\epsilon m_t} = \delta_t$. Thus, by the union bound,

   \[
   \mathbb{P}\left[ \exists t \geq 1: h_t \text{ consistent}\mid R(h_t) \geq \epsilon \right] \leq \sum_{t=1}^{+\infty} \delta_t = \frac{\delta}{2} \sum_{t=1}^{+\infty} \frac{1}{t^2} = \frac{\delta \pi^2}{2} \frac{1}{6} \leq \delta.
   \]
Let \( t^* \) be the time at which the algorithm terminates. \( t^* \) is upper bounded by the index \( t \) such that \( h_t = f^* \). If we reuse samples, at most \( \frac{1}{\sqrt{2\pi^2}} \) points are needed overall.

(c) Can you generalize these results to the case where \( \mathcal{H} = \bigcup_{n=1}^{+\infty} \mathcal{H}_n \) with \( \text{VCdim}(\mathcal{H}_n) = d_n < +\infty \)?

Solution: Same algorithm, except at round \( t \) a consistent hypothesis in \( \mathcal{H}_t \) is sought. Assume that the ordering of \( \mathcal{H}_n \) is such that \( \mathcal{H}_n \subseteq \mathcal{H}_{n+1} \). At each round \( t \), select a sample \( S_{m_t} \) of size \( m_t \) and return \( h_t \in \mathcal{H}_t \) if it is consistent with \( S_{m_t} \). To derive the error bound, let \( \delta_t = \frac{\delta}{2^t} \) and let \( m_t = O\left(\frac{d_t \log 1}{\delta_t} \right) \) and observe that:

\[
P \left( R_D(h_t) > \epsilon \right) \leq P \left( \cup_{t=0}^{\infty} \{ \exists h \in \mathcal{H}_t : \hat{R}_{S_{m_t}}(h) = 0, R_D(h) > 0 \} \right) \leq \sum_{t=1}^{\infty} \delta_t = \frac{\delta}{2} \sum_{t=1}^{\infty} \frac{1}{t^2} \leq \delta.
\]

2. Suppose \( S \) is an infinite set that can be fully shattered by \( \mathcal{H} \). We wish to show that \( \mathcal{H} \) cannot be written as a countable union \( \mathcal{H} = \bigcup_{n=1}^{+\infty} \mathcal{H}_n \) with \( \text{VCdim}(\mathcal{H}_n) = d_n < +\infty \).

(a) Show that we can define a family of subsets \((X_n)_{n \geq 1}\) such that \(|X_n| = d_n + 1\) and \(X_n \subseteq S - \bigcup_{1 \leq k \leq n-1} X_k\).

Solution: This is straightforward since \( S \) is an infinite sample and since \( d_n \) is finite for any \( n \geq 1 \).

(b) Show that for any \( n \geq 1 \), there exists a labeling \( X_n^t \) that cannot be obtained using \( \mathcal{H}_n \).

Solution: This follows directly the definition of the VC-dimension: no set of size \( d_n + 1 \) can be fully shattered by \( \mathcal{H}_n \).

(c) Consider the labeling \( X^t \) of \( X = \bigcup_{n=1}^{+\infty} X_n \) obtained using all the \( X_n^t \)s. Show that no labeling of \( S \) using \( \mathcal{H} \) can be consistent with \( X^t \). Conclude that that \( \mathcal{H} \) cannot be written as a countable union \( \mathcal{H} = \bigcup_{n=1}^{+\infty} \mathcal{H}_n \) with \( \text{VCdim}(\mathcal{H}_n) = d_n < +\infty \).
Solution: Note that, by definition, all $X_n$'s are disjoint. Thus, the labeling $X^l$ obtained from all $X_n^l$'s is well defined. Let $Y$ be a labeling of $T$ consistent with $X^l$. Then, for any $n \geq 1$, $Y_{|X_n}$ is a labeling of $X_n$ matching $X_n^l$ and thus $Y$ is not in $\mathcal{H}_n$. Since $Y$ is not in $\mathcal{H}_n$ for any $n \geq 1$, it is not in $\mathcal{H}$. This shows that the assumption that $\mathcal{H}$ cannot be written as a countable union $\mathcal{H} = \bigcup_{n=1}^{+\infty} \mathcal{H}_n$ with $\text{VCdim}(\mathcal{H}_n) = d_n < +\infty$ does not hold.

3. Suppose you only know an upper bound $\alpha_n$ on $\text{VCdim}(\mathcal{H}_n) = d_n < +\infty$ with $\sum_{n=1}^{+\infty} e^{-\alpha_n} < +\infty$. Give a generalization bound for the SRM-type algorithm defined by

$$f^* = \arg\min_{k \geq 1, h \in \mathcal{H}_k} \hat{R}_S + \sqrt{\frac{32\alpha_k \log(em)}{m}}.$$ 

for a sample $S$ of size $m$.

Solution: Let $F_k(h) = \hat{R}_S + \sqrt{\frac{32\alpha_k \log(em)}{m}}$. Then using $\mathcal{H} = \bigcup_{k=1}^{+\infty} \mathcal{H}_k$

$$\mathbb{P}\left( \sup_{h \in \mathcal{H}} R(h) - F_k(h)(h) - \sqrt{\frac{2dk(h) \log(em)/d_k(h)}{m}} > \epsilon \right)$$

can be bounded as follows:

$$\leq \sum_{k=1}^{+\infty} \mathbb{P}\left( \sup_{h \in \mathcal{H}_k} R(h) - F_k(h) - \sqrt{\frac{2dk \log(em)/d_k}{m}} > \epsilon \right)$$

$$= \sum_{k=1}^{+\infty} \mathbb{P}\left( \sup_{h \in \mathcal{H}_k} R(h) - F_k(h) - \hat{R}_S(h) - \sqrt{\frac{2dk \log(em)/d_k}{m}} > \epsilon + \sqrt{\frac{32\alpha_k \log(em)}{m}} \right)$$

$$\leq \sum_{k=1}^{+\infty} \exp\left( - 2m \left( \epsilon + \sqrt{\frac{32\alpha_k \log(em)}{m}} \right)^2 \right)$$

$$\leq \sum_{k=1}^{+\infty} \exp\left( - 2me^2 \right) \exp\left( - a_k \log m \right)$$

$$\leq Ce^{-2me^2}.$$ 

Applying similar steps and recalling that $f^*$ is the minimizer of $\hat{R}_S +$
\[ \sqrt{\frac{32\alpha_k \log(em)}{m}}, \] we can show that

\[ \mathbb{P}\left( \sup_{h \in \mathcal{H}} F_k(f^*)(f^*) - R(h^*) - \sqrt{\frac{32\alpha_k (h^*) \log(em)}{m}} - \sqrt{\frac{2dk(h^*) \log em/ dk(h^*)}{m}} > \frac{\epsilon}{2} \right) \leq e^{-\frac{m\epsilon^2}{2}}. \]

Combining the results above and the union bound provides the generalization bound with \( \delta = (1 + C)e^{-\frac{m\epsilon^2}{2}}. \)

B. Learning kernels

Let \( \mathcal{K} \) be the family of all Gaussian kernels defined over \( \mathbb{R}^N \):

\[ \mathcal{K} = \left\{ K_\gamma : K_\gamma(x, x') = e^{-\gamma \|x-x'\|^2}, \forall x, x' \in \mathbb{R}^N, \gamma > 0 \right\}. \]

Consider the hypothesis set defined via the reproducing kernel Hilbert space of the kernels in \( \mathcal{K} \):

\[ \mathcal{H} = \left\{ h : h \in \mathbb{H}_K, K \in \mathcal{K}, \|h\|_{\mathbb{H}_K} \leq 1 \right\}. \]

1. Let \( S = (x_1, \ldots, x_m) \) be a sample of size \( m \). Show that \( \hat{\mathcal{R}}_S(\mathcal{H}) = \frac{1}{m} \mathbb{E}_\sigma \left[ \sqrt{\sup_{\gamma > 0} \sigma^T K_\gamma \sigma} \right] \), where \( K_\gamma \) is the Gram matrix of kernel \( K_\gamma \) for the sample \( S \).
Solution:

\[
\hat{R}_S(\mathcal{H}) = \frac{1}{m} \mathbb{E} \left[ \sup_{h \in \mathcal{H}_K, \|h\|_k \leq 1} \sum_{i=1}^{m} \sigma_i \langle h, \Phi_K(x_i) \rangle \right]
\]

\[
= \frac{1}{m} \mathbb{E} \left[ \sup_{K \in \mathcal{K}} \left\| \sum_{i=1}^{m} \sigma_i \Phi_K(x_i) \right\|_{\mathbb{H}_K} \right]
\]

\[
= \frac{1}{m} \mathbb{E} \left[ \sup_{K \in \mathcal{K}} \left( \sum_{i=1}^{m} \sigma_i \Phi_K(x_i) \right)^2 \right]
\]

\[
= \frac{1}{m} \mathbb{E} \left[ \sup_{\gamma > 0} \sqrt{\sigma^\top K_{\gamma} \sigma} \right]
\]

\[
= \frac{1}{m} \mathbb{E} \left[ \sqrt{\sup_{\gamma > 0} \sigma^\top K_{\gamma} \sigma} \right].
\]

2. Suppose \(\|x_i - x_j\| = 1\) for \(i \neq j\). Compute exactly \(\hat{R}_S(\mathcal{H})\).

Solution: Given that \(\|x_i - x_j\| = 1\) for \(i \neq j\), the diagonal terms of the kernel matrix are \(K_{i,j}^{i,j} = 1\) for \(i = j\) and the off-diagonal terms are \(K_{i,j}^{i,j} = e^{-\gamma}\) for \(i \neq j\).

\[
\sup_{\gamma > 0} \left[ \sigma^\top K_{\gamma} \sigma \right] = \sup_{\gamma > 0} \left[ \sum_{i,j} \sigma_i \sigma_j K_{i,j}^{i,j} \right]
\]

\[
= \sup_{\gamma > 0} \left[ m + e^{-\gamma} \sum_{i \neq j} \sigma_i \sigma_j \right]
\]

\[
= \sup_{\gamma > 0} \left[ \sum_{i,j} \sigma_i \sigma_j K_{i,j}^{i,j} \right]
\]

\[
= m + \sup_{\gamma > 0} e^{-\gamma} \sum_{i \neq j} \sigma_i \sigma_j
\]

\[
= m + \sum_{i \neq j} \sigma_i \sigma_j \mathbb{1}_{\sum_{i \neq j} \sigma_i \sigma_j > 0}.
\]
Observe that:

\[ m + \sum_{i \neq j} \sigma_i \sigma_j = \sum_{i,j=1}^{m} \sigma_i \sigma_j = \sigma^\top 11^\top \sigma = (\sigma^\top 1)^2 = \left[ \sum_{i=1}^{m} \sigma_i \right]^2. \]

It is also known that:

\[
\mathbb{E} \left[ \left| \sum_{i=1}^{m} \sigma_i \right| \right] = \frac{1}{2^{m-1}} \left\lceil \frac{m}{2} \right\rceil \left( \left\lfloor \frac{m}{2} \right\rfloor \right) \leq \sqrt{m}. \quad \text{(Jensen’s ineq.)}
\]

Thus, we have:

\[
\sup_{\gamma > 0} [\sigma^\top K_{\gamma} \sigma] = \begin{cases} 
|\sum_{i=1}^{m} \sigma_i| & \text{if } \sum_{i \neq j} \sigma_i \sigma_j > 0; \\
\sqrt{m} & \text{if } \sum_{i \neq j} \sigma_i \sigma_j < 0; \\
\sqrt{m} & \text{if } \sum_{i \neq j} \sigma_i \sigma_j = 0.
\end{cases}
\]

When \( m \) is odd, the event \( \sum_{i \neq j} \sigma_i \sigma_j = 0 \) cannot occur and the other two events are symmetric, each with probability 1/2. Thus, we have:

\[
\widehat{R}_S(\mathcal{H}) = \frac{1}{2m} \left( \frac{m}{2} + 1 \right) + \frac{1}{2} \sqrt{m}.
\]

When \( m \) is even, the event \( \sum_{i \neq j} \sigma_i \sigma_j = 0 \) occurs with probability \( \frac{1}{2^m} \left( \left\lceil \frac{m}{2} \right\rceil \right) \) and the other two events with equal probability \( p = \frac{1}{2} - \frac{1}{2^{m+1}} \left( \left\lceil \frac{m}{2} \right\rceil \right) \). Thus, we have:

\[
\widehat{R}_S(\mathcal{H}) = \left[ \frac{1}{2} - \frac{1}{2^m+1} \left( \left\lceil \frac{m}{2} \right\rceil \right) \right] \frac{1}{2^m} \left( \frac{m}{2} \right) + \frac{1}{2} \left[ \frac{1}{2} + \frac{1}{2^m+1} \left( \left\lceil \frac{m}{2} \right\rceil \right) \right] \frac{1}{\sqrt{m}}.
\]

We can express the solution in terms of \( \beta_0 \approx \sqrt{\frac{2}{\pi}} \), where \( \frac{1}{m} \mathbb{E}[|\sum_{i=1}^{m} \sigma_i|] = \frac{\beta_0}{\sqrt{m}} \), as follows:

\[
\widehat{R}_S(\mathcal{H}) = \begin{cases} 
\frac{1}{2} \left[ \beta_0 + 1 \right] \frac{1}{\sqrt{m}} & \text{if } m \text{ even} \\
\frac{1}{2} \left[ \beta_0 + 1 \right] \frac{1}{\sqrt{m}} + \frac{1}{2} \left[ \beta_0 - \beta_0^2 \right] \frac{1}{m} & \text{otherwise}.
\end{cases}
\]