1. [15 points] As shown in the course notes, if $\mathcal{F}_\lambda$ is the zero centered ball in the RKHS generated by a Mercer kernel $K$, then for independent and identically distributed (iid) samples $X^n = (X_1, \ldots, X_n)$,

$$E_{X^n} R_n(\mathcal{F}_\lambda(X^n)) \leq \lambda \sqrt{\frac{E K(X, X)}{n}}.$$ 

Here is a different proof based on the representation, $K(x, x') = \sum_j \psi_j(x) \psi_j(x')$, of $K$ in terms of a complete orthonormal basis $(\psi_i)$ for the RKHS. By definition,

$$\sup_{f \in \mathcal{F}_\lambda} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i) \right| = \frac{1}{n} \sup_{c:||c|| \leq \lambda} \left| \sum_{i=1}^n \sigma_i c_j \psi_j(X_i) \right|$$

$$= \frac{1}{n} \sup_{c:||c|| \leq \lambda} \left| \sum_j c_j \left( \sum_{i=1}^n \sigma_i \psi_j(X_i) \right) \right|$$

$$= (a) \frac{\lambda}{n} \sqrt{\sum_j \left( \sum_{i=1}^n \sigma_i \psi_j(X_i) \right)^2}$$

Thus,

$$E_{X^n} R_n(\mathcal{F}_\lambda(X^n)) = \frac{\lambda}{n} E_{X^n} E_\sigma \left[ \sum_j \left( \sum_{i=1}^n \sigma_i \psi_j(X_i) \right)^2 \right]$$

$$\leq (b) \frac{\lambda}{n} \sqrt{E_{X^n} \sum_j \left( \sum_{i=1}^n \sigma_i \psi_j(X_i) \right)^2}$$

$$\leq (c) \frac{\lambda}{n} \sqrt{E_{X^n} \sum_j \psi_j(X_i)^2}$$

$$= \frac{\lambda}{n} \sqrt{\sum_{i=1}^n E_{X^n} K(X_i, X_i)} = \frac{\lambda}{n} \sqrt{\sum_{i=1}^n E K(X_i, X)} = \lambda \sqrt{\frac{E K(X, X)}{n}}$$

Explain what properties are used for deriving the steps labeled by (a), (b), and (c).

Solution: (a) is by the Cauchy-Schwarz inequality, which holds with equality when vectors are proportional. That is, for any vector $v \in \ell^2$, $||c|| = ||c||$, with equality if $c$ is a scaler multiple of $v$.

(b) follows by Jensen’s inequality applied for the concave function $\sqrt{x}$. That is, $E \left[ \sqrt{Y} \right] \leq \sqrt{E[Y]}$ for any random variable $Y$.

(c) follows since the $\sigma_i$’s are independent, mean zero, variance one random variables. So for a vector $v$, $E_\sigma \left[ (\sum_i \sigma_i v_i)^2 \right] = E_\sigma \sum_i \sum_{i'} \sigma_i \sigma_{i'} v_i v_{i'} = \sum_i \sum_{i'} 1_{(i=i')} v_i v_{i'} = \sum_i (v_i)^2$.

2. [16 points] Suppose $X_1, \ldots, X_n$ are independent and identically distributed random vectors in $\mathbb{R}^d$ for some positive integers $n$ and $d$. Given $X_1, \ldots, X_n$, let $K_n$ denote the minimum number of balls of radius five in $\mathbb{R}^d$ that cover all $n$ values $X_1, \ldots, X_n$. Find a concentration inequality for $K_n$ (i.e. an upper bound on the probability of an event of the form $\{|K_n - E[K_n]| \geq t\}$.) Justify your answer.
4. \[ K_n \text{ is a function } F(X_1, \ldots, X_n). \text{ The function } F \text{ has the bounded differences property for constants } c_1 = \cdots = c_n = 1. \text{ Indeed, if one argument of } F(x_1, \ldots, x_n) \text{ is changed, the most the covering number could increase is one, because one could take a covering of the original } n \text{ points and add one more ball to cover the moved point. It follows that } F \text{ can decrease by at most one as well. Thus, by McDiarmid’s inequality,} \]

\[ P \{|K_n - \mathbb{E}[K_n]| \geq t\} \leq 2 \exp \left( -\frac{2t^2}{n} \right). \]

Equivalently, \[ P \{|K_n - \mathbb{E}[K_n]| \geq \alpha \sqrt{n}\} \leq 2 \exp(-2\alpha^2). \]

3. \textbf{[16 points]} It is shown by a direct proof at the end of the supplementary notes for the first two weeks that if a concept learning problem \((X, \mathcal{P}, \mathcal{C})\) is such that \(|\mathcal{C}| = M \) (i.e. there are only \(M\) concepts) then \(\mathcal{C}\) is PAC learnable and the ERM algorithm for 0-1 loss is such that \(P^n\{L_P(\hat{C}_n) > L_P(C) + \epsilon\} \leq \delta\) whenever \(2Me^{-nc/2} \leq \delta\). Equivalently, with probability at least \(1 - \delta\),

\[ L_P(\hat{C}_n) \leq L_P(C) + \sqrt{\frac{2}{n} \log \left( \frac{2M}{\delta} \right)}. \]

Explain how the PAC learnability in this case also follows from the more general Fundamental Theorem of Concept Learning, and find what performance upper bound is provided by following that method.

\textbf{Solution:} The VC dimension of \(\mathcal{C}\) is bounded above by \(V(\mathcal{C}) \leq \log_2 M\) because for any \(n\) points, \(|\mathcal{C} \cap \mathcal{C} \subseteq \mathcal{C}| \leq |\mathcal{C}| = M\), so if \(2^n > M\) then \(n\) points cannot be shattered by \(\mathcal{C}\). Therefore, the fundamental theorem of concept learning gives that with probability at least \(1 - \delta\),

\[ L_P(\hat{C}_n) \leq L_P(C) + 8 \sqrt{\left(\log_2 M\right)\left(\log(n+1)\right)} + \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}}. \tag{1} \]

Also, for a universal constant \(C\),

\[ L_P(\hat{C}_n) \leq L_P(C) + C \sqrt{\frac{\log_2 M}{n}} + \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}}. \tag{2} \]

4. \textbf{[20 points]} Let \(X = \ell^2\), where \(\ell^2\) is the Hilbert space of sequences \(\ell^2 = \{(x_1, x_2, \ldots) : \sum x_i^2 < \infty\}\), with the inner product \(\langle x, x' \rangle = \sum x_i x'_i\). Consider the concept learning problem \((X, Y = \{\pm 1\}, \mathcal{P}_X, \mathcal{H})\) in the realizable case, where \(\mathcal{P}_X\) is the set of probability measures on \(X\), and \(\mathcal{H}\) is the set of half-space classifiers of the form \(h(x) = \text{sgn}(\langle w, x \rangle)\), where \(w \in \ell^2\).

(a) \textbf{(10 points)} Show that this problem is not PAC learnable.

\textbf{Solution:} (Proofs can vary, but in one way or another, this is a form of no free lunch theorem.)

Let \(x^{(i)}\) be the \(i^{th}\) unit vector in \(X\), so \(x^{(i)}_j = 1_{\{i=j\}}\). For a classifier \(h(x) = \text{sgn}(\langle w, x \rangle)\), note that \(h(x^{(i)}) = \text{sgn}(w_i)\). Thus, for any \(n \geq 1\), the labels of \(x^{(1)}, \ldots, x^{(n)}\) can be given by any \(\{\pm 1\}\) vector of length \(n\), by choice of \(w\). That is, \(x^{(1)}, \ldots, x^{(n)}\) is shattered by \(\mathcal{H}\) for any \(n \geq 1\), so the VC dimension of \(\mathcal{H}\) is infinite. It follows from the converse half (aka no free lunch theorem half) of the fundamental theorem of concept learning, that the learning problem is not PAC learnable. Another proof is to note that \(\mathcal{H}\) contains a Dudley class of dimension \(n\) for any \(n \geq 1\). Therefore, the VC dimension is infinite, so by the fundamental theorem of concept learning the problem is not PAC.

(b) \textbf{(10 points)} Does the problem become PAC learnable if \(\mathcal{H}\) is replaced by \(\mathcal{H}_\lambda\) for some fixed \(\lambda > 0\), where \(\mathcal{H}_\lambda = \{h(x) = \text{sgn}(\langle w, x \rangle) : \|w\| \leq \lambda\}\)?> Explain.

\textbf{Solution:} No, even with that constraint, the problem still has infinite VC dimension. For any \(n \geq 1\), the labels of \(x^{(1)}, \ldots, x^{(n)}\) can be any vector in \(\{\pm\}^n\) even with \(\|w\| \leq \lambda\). For example, let \(w_i = \pm \lambda \sqrt{-1}^i\) for each \(i \geq 1\). Then \(\|w\| = \lambda\) and \(h(x^{(i)}) = \text{sgn}(w_i)\) for all \(i\).

The alternative solution based on Dudley class mentioned for part (a) continues to hold as well. For any \(w \in \ell^2\), \(w\) and the scaled version \(\frac{\lambda}{\|w\|} w\) correspond to the same classifier \(h\). That is, \(\mathcal{H} = \mathcal{H}_\lambda\) for any \(\lambda > 0\).
Remark: This negative result does not contradict the positive results we have seen for classification using a surrogate loss function. The bounds imply PAC learnability under the original loss function only if there is some condition restricting how many samples can be in the margin, such as in problem 2 of problem set 4.

5. [17 points] Suppose $C_1$ and $C_2$ are families of subsets of some feature space $X$. Suppose both $C_1$ and $C_2$ are VC families (i.e. have finite VC dimension). Is $C_1 \cup C_2$ also a VC family? If so, explain how $V(C_1 \cup C_2)$ can be bounded above in terms of $V(C_1)$ and $V(C_2)$. If not, give a counter example. (Hint: The Sauer-Shelah lemma is relevant.)

**Solution:** Yes. Let $V(C_i) = d_i$. We are given $d_1$ and $d_2$ are finite. Let $S = \{x_1, \ldots, x_n\} \subseteq X$. By the Sauer-Shelah lemma, $|\{S \cap C : C \in C_i\}| \leq (n+1)^{d_i}$ for $i = 1$ and $i = 2$. Therefore, $|\{S \cap C : C \in C_1 \cup C_2\}| \leq (n+1)^{d_1} + (n+1)^{d_2}$. Thus, if $n$ is large enough that $(n+1)^{d_1} + (n+1)^{d_2} < 2^n$ then $C_1 \cup C_2$ cannot shatter any set of $n$ points. So $V(C_1 \cup C_2) \leq \min\{n \geq 1 : (n+1)^{d_1} + (n+1)^{d_2} < 2^n\} - 1$.

Better yet, we know that $|\{S \cap C : C \in C_1 \cup C_2\}| \leq \binom{n}{\leq d_1} + \binom{n}{\leq d_2} = \binom{n}{d_1} + \binom{n}{d_2} = 2^n$ for $n = d_1 + d_2 + 1$ and $\binom{n}{\leq d_1} + \binom{n}{\leq d_2} < 2^n$ for $n \geq d_1 + d_2 + 2$. Therefore, $V(C_1 \cup C_2) \leq V(C_1) + V(C_2) + 1$.

This bound is tight. For example, consider the case $n = d_1 + d_2 + 1$, $C_1 = \{[n]_{\leq d_1}\}$ (all subsets of $[n]$ with cardinality less than or equal to $d_1$), and $C_2 = \{[n]_{\leq d_2}\}$.

6. [16 points] Suppose that $X = \{1, \ldots, n\} = [n]$. Functions on $X$ are equivalent to column vectors, and the space of all functions on $X$ is just $\mathbb{R}^n$. Let $K = (K_{i,j})_{i,j \in [n]}$ be a symmetric, positive definite matrix (so $K$ is full rank). Then $K$ is a Mercer kernel and the set of vectors in $\mathcal{H}_K$, the column span of $K$, is $\mathbb{R}^n$. Let $e_1, \ldots, e_n$ denote the standard orthonormal basis for $\mathbb{R}^n$, so that $e_i$ is the column vector with $i^{th}$ coordinate equal to one and zeros elsewhere. Let $Q_{i,j} = \langle e_i, e_j \rangle_K$ for $i, j \in [n]$.

(a) (4 points) Show that $\langle x, x' \rangle_K = \sum_{j,j'} x_j Q_{j,j'} x'_j$ (i.e. $\langle x, x' \rangle_K = x^T Q x'$) for all $x, x' \in \mathbb{R}^n$.

**Solution:** For any $x, x' \in \mathbb{R}$,

$$\langle x, x' \rangle_K = \left( \sum_{i=1}^n x_i e_i, \sum_{j=1}^n x_j' e_j \right)_K = \sum_{i=1}^n \sum_{j=1}^n x_i x_j' \langle e_i, e_j \rangle_K = x^T Q x'.$$

(b) (12 points) Express the matrix $Q$ in terms of the matrix $K$.

**Solution:** By the definition of $\langle \cdot, \cdot \rangle_K$, $(K, K_{i'})_K = K_{i,i'}$, where $K_i$ is the $i^{th}$ column of $K$. Using part (a) yields $K_{i,i'} = \langle K_{i} K_{i'} \rangle_K = \sum_{j,j'} K_{i,j} Q_{j,j'} K_{i',j'} = \sum_{j,j'} K_{i,j} Q_{j,j'} K_{j',i'} = (K Q K)_{i,i'}$. That is, $K = K Q K$. Multiplying each side by $K^{-1}$ yields $I = K Q$. Therefore, $Q = K^{-1}$.

Another solution is to note that $1_{\{i=j\}} = \langle K_i, e_j \rangle_K = K_i^T Q e_j$, which in matrix form is $I = K^T Q$. Since $K$ is symmetric and full rank, it follows that $Q = K^{-1}$.