Solutions to Graded Homework 1 (HW3)
CS-526 Learning Theory

Problem 1. Exercise 5 of Chapter 3
Consider some $h \in \mathcal{H}$ s.t. $L_{(\overline{\mathcal{D}_m},f)}(h) > \epsilon$. Then it follows by definition of $L_{(\overline{\mathcal{D}_m},f)}$ (the true loss) that

$$\frac{\sum_{i=1}^{m} \mathbb{P}_{X\sim\mathcal{D}_i}[h(X) = f(X)]}{m} < 1 - \epsilon.$$

Then we bound the probability that $L_S(h) = 0$ as follows:

$$\begin{aligned}
\mathbb{P}_{S\sim\prod_{i=1}^{m}\mathcal{D}_i}[L_S(h) = 0] &= \prod_{i=1}^{m} \mathbb{P}_{X\sim\mathcal{D}_i}[h(X) = f(X)] \\
&= \left(\left(\prod_{i=1}^{m} \mathbb{P}_{X\sim\mathcal{D}_i}[h(X) = f(X)]\right)^{1/m}\right)^{m} \\
&\leq \left(\frac{\sum_{i=1}^{m} \mathbb{P}_{X\sim\mathcal{D}_i}[h(X) = f(X)]}{m}\right)^{m} \\
&\leq (1-\epsilon)^m \\
&\leq e^{-\epsilon m}.
\end{aligned}$$

The first inequality above is the geometric-arithmetic mean inequality. It remains to apply the union bound to conclude that the probability that there exists $h \in \mathcal{H}$ with true loss greater that ϵ and zero loss for the observed samples is upper bounded by $|\mathcal{H}|e^{-\epsilon m}$.

Problem 2. Exercise 2 of Chapter 5
Denote by $\mathcal{H}_d$ the class of axis-aligned rectangles in $\mathbb{R}^d$. Since $\mathcal{H}_2 \subseteq \mathcal{H}_5$, the latter class is more representative and in the presence of sufficiently many samples is likely to have smaller true loss. However, the growth rate (and the VC dimension) of the former is smaller, which gives us better bounds on the approximation error, which makes it more preferable in case of small number of samples.

Problem 3. Exercise 3 of Chapter 5 We modify the proof of NFL theorem in the book. Consider subset $C \subset \mathcal{X}$ of size km. There are $T = 2^{km}$ functions from C to $\{0,1\}$, $f_1, \ldots, f_T$. Let D_i's be probability distribution defined as in the proof of NFL.

Fix a training set $S_j = \{x_1, \ldots, x_m\}$ and let $v_1, \ldots, v_p$ be the examples in C that do not appear in S_j. Clearly, we have $p \geq (k-1)m$. Therefore, for every $h : C \to \{0,1\}$ and $i \in [T]$,

we have (Eq. 5.5 in the book)

$$\begin{aligned} L_{D_i}(h) &= \frac{1}{km}\sum_{x\in C} \mathbb{1}_{[h(x)\neq f_i(x)]} \\ &\geq \frac{1}{km}\sum_{r=1}^{p} \mathbb{1}_{[h(v_r)\neq f_i(v_r)]} \\ &\geq \frac{k-1}{kp}\sum_{r=1}^{p} \mathbb{1}_{[h(v_r)\neq f_i(v_r)]} \end{aligned} \tag{1}$$

Then, the right-hand side of Eq. 5.6 becomes

$$\frac{k-1}{k}\min_{r\in[p]}\frac{1}{T}\sum_{i=1}^{T} \mathbb{1}_{[A(S_j^i)(v_r)\neq f_i(v_r)]}$$

From which we can conclude (similar to the proof of NFL) that

$$\mathbb{E}\Big[L_D\big(A(S)\big)\Big] \geq \frac{k-1}{2k}$$

Problem 4. Exercise 2 of Chapter 6

1. We show that $\text{VCdim}(\mathcal{H}_{=k}) = \min(k, |\mathcal{X}| - k)$.

Consider a set of $k+1$ elements. The all-one labeling cannot be obtained, hence $\text{VCdim}(\mathcal{H}) \leq k$. Analogously, for a set of $|\mathcal{X}| - k + 1$ elements, the all-zero labeling cannot be obtained so $\text{VCdim}(\mathcal{H}_{=k}) \leq \min(k, |\mathcal{X}| - k)$.

Take a set C of size $m = \min(k, |\mathcal{X}| - k)$ and a labeling $(y_1, \ldots, y_m)$ with s ones, $0 \leq s \leq m$. We can pick a hypothesis $h \in \mathcal{H}_{=k}$ such that $h(x_i) = y_i$ for all $x_i \in C$ and it has $k - s$ ones on the set $\mathcal{X} \setminus C$ (this is possible because $|\mathcal{X} \setminus C| \geq k \geq k - s$). Therefore, C is shattered and $\text{VCdim}(\mathcal{H}_{=k}) \geq \min(k, |\mathcal{X}| - k)$.

2. We show that $\text{VCdim}(\mathcal{H}_{at-most-k}) = \min(2k+1, |\mathcal{X}|)$.

Consider a set of $2k+2$ elements. It is clear that any labeling with $k+1$ ones and $k+1$ zeros cannot be obtained, so $\text{VCdim}(\mathcal{H}_{at-most-k}) \leq 2k+1$. Note that it may happen that $2k+1 > |\mathcal{X}|$, so the bound should be $\text{VCdim}(\mathcal{H}_{at-most-k}) \leq \min(2k+1, |\mathcal{X}|)$.

Take a set of $\min(2k+1, |\mathcal{X}|)$ elements. Any labeling on this set has either $\leq k$ zeros or $\leq k$ ones, so it is shattered by $\mathcal{H}_{at-most-k}$. Thus, $\text{VCdim}(\mathcal{H}_{at-most-k}) \geq \min(2k+1, |\mathcal{X}|)$.

Problem 5. Exercise 8 of Chapter 6

Let's first prove the lemma. Let $m \geq 1$ and $0.x_1x_2\ldots$ the binary expansion of $x \in (0,1)$. Assume that there exists $k \geq m$ such that $x_k = 1$. We have:

$$\begin{aligned} \sin(2^m\pi x) = \sin(2^m\pi \cdot (0.x_1x_2\ldots)) &= \sin(2\pi \cdot (x_1x_2\ldots x_{m-1}.\, x_m x_{m+1}\ldots)) \\ &= \sin(2\pi \cdot (0.x_m x_{m+1}\ldots))\,. \end{aligned}$$

If $x_m = 0$ then $\exists k > m$ s.t. $x_k = 1$, i.e., the number $0.0x_{m+1}\ldots$ is nonzero. This means that $2\pi \cdot (0.0x_{m+1}\ldots) \in (0, \pi)$ where $\sin(x)$ is positive, which gives the label 1. If $x_m = 1$ then $2\pi \cdot (0.1x_{m+1}\ldots) \in [\pi, 2\pi)$ where $\sin(x)$ is non positive, which gives the label 0. This ends the proof of the lemma.

To prove that $\mathcal{H}$ has infinite VC-dimension, we need to show that for any n there is a set of n points in $\mathbb{R}$ on which we can obtain all 2^n possible labelings. Let $(x_1, \ldots, x_n) \in (0,1)^n$ such that the first 2^n bits of their binary expansions give all possible labelings and their $(2^n)^{\text{th}}$ bit is always one (i.e., the $(2^n)^{\text{th}}$ bits form the all-one labeling).

Example for $n = 3$:

x_1	0.	0	1	0	1	0	1	0	1
x_2	0.	0	0	1	1	0	0	1	1
x_3	0.	0	0	0	0	1	1	1	1

Using the lemma, invoking the function $\lceil \sin(2^i \pi x) \rceil$ on the set $\{x_1, \ldots, x_n\}$ for $1 \leq i \leq 2^n$ allows to obtain all possible labelings. Hence, $\mathcal{H}$ shatters the set $\{x_1, \ldots, x_n\}$.

Problem 6. Stable learning

1. Fix the labeling function f and a distribution $\mathcal{D}$ on $\mathcal{X}$. Call a hypothesis $h \in \mathcal{H}$ "bad" if $P_{x \sim \mathcal{D}}\big[h(x) \neq f(x)\big] > \epsilon$. Let E_h be the event that m independent samples in S drawn from $\mathcal{D}$ are all consistent with h, i.e. $h(x_i) = f(x_i)$, for $1 \leq i \leq m$. Then, if h is bad, $P[E_h] \leq (1-\epsilon)^m \leq e^{-\epsilon m}$.

   Consider the event

   $$E = \bigcup_{\text{bad } h \in \mathcal{H}} E_h$$

   Then, by union bound, we have:

   $$P[E] \leq \sum_{\text{bad } h \in \mathcal{H}} P[E_h] \leq |\mathcal{H}| e^{-\epsilon m}$$

   If $m \geq \frac{1}{\epsilon}\big(\log|\mathcal{H}| + \log\frac{1}{\delta}\big)$, then this probability is upper bounded by δ.

   Thus, whenever m is larger than the bound, the probability that a stable learner returns a bad hypothesis $h_S \in E$ is at most δ. Which means that the event $P(h_S(x) \neq f(x)) > \epsilon$ has probability at most δ. Thus the event $P(h_S(x) \neq f(x)) > \epsilon$ has probability at least $1 - \delta$.

2. (a) The output is $h = z_1 \wedge \bar{z}_2 \wedge \bar{z}_3 \wedge z_4$. Stability is checked by plugging all four $x_i \in S$ and checking that $h(x_i) = \phi^*(x_i)$.

   (b) We have that $|\mathcal{H}| = 3^n$, because any variable can appear as z_i or $\bar{z}_i$, or do not appear in a conjunction. Then using part 1, we should have

   $$m \geq \frac{1}{\epsilon}(\log|\mathcal{H}| + \log\frac{1}{\delta}) = \frac{1}{\epsilon}(n\log 3 + \log\frac{1}{\delta})$$