Solution 1: 19 February 2019
CS-526 Learning Theory

Exercise 3.1

The hypothesis class $\mathcal{H}$ being PAC learnable with sample complexity $m_{\mathcal{H}}(\cdot,\cdot)$ means that there is a learning algorithm A such that when running A on $m \geq m_{\mathcal{H}}(\epsilon,\delta)$ i.i.d. samples generated by $\mathcal{D}$ and labeled by f, with probability at least $1-\delta$, A returns a hypothesis $h \in \mathcal{H}$ with $L_{D,f}(h) \leq \epsilon$.

Given $0 < \epsilon_1 \leq \epsilon_2 < 1$, consider $m \geq m_{\mathcal{H}}(\epsilon_1,\delta)$, we have that with probability at least $1-\delta$, A returns a hypothesis $h \in \mathcal{H}$ with $L_{D,f}(h) \leq \epsilon_1 \leq \epsilon_2$. This implies that $m_{\mathcal{H}}(\epsilon_1,\delta)$ is a sufficient number of samples for accuracy ϵ_2. Therefore, $m_{\mathcal{H}}(\epsilon_1,\delta) \geq m_{\mathcal{H}}(\epsilon_2,\delta)$.

The proof of $m_{\mathcal{H}}(\epsilon,\delta_1) \geq m_{\mathcal{H}}(\epsilon,\delta_2)$ for $0 < \delta_1 \leq \delta_2 < 1$ follows analogously from the definition.

Exercise 3.3

The realizability assumption for $\mathcal{H} = \{h_r : r \in \mathbb{R}_+\}$ implies that there is a circle such that any x inside it has label $y = 1$, and the learning task here is to distinguish this circle. Now consider an ERM algorithm which given a training sequence $S = \{(x_i, y_i)\}_{i=1}^m$, returns the hypothesis $\hat{h}$ corresponding to the tightest circle which contains all the positive instances in S where $y_i = 1$ and does not allow false negative predictions. With the realizability assumption let h^* be the circle with zero training error and r^* be the corresponding radius.

Let $\bar{r} \leq r^*$ be a scalar such that $\mathbb{P}_{x\sim\mathcal{D}}(x : \bar{r} \leq \|x\| \leq r^*) = \epsilon$ and $E = \{x \in \mathbb{R}^2 : \bar{r} \leq \|x\| \leq r^*\}$. We have

$$
\begin{aligned}
\mathbb{P}(L_{\mathcal{D}}(h_S) \geq \epsilon) &\leq \mathbb{P}(\text{no points in } S \text{ belongs to } E) \\
&= (1-\epsilon)^m \\
&\leq e^{-\epsilon m}
\end{aligned}
$$

The desired bound on the sample complexity follows from requiring $e^{-\epsilon m} \leq \delta$.

Exercise 3.7

Let g be any (potentially probabilistic) classifier from $\mathcal{X}$ to $\{0, 1\}$. Note that for 0-1 loss

$$
\begin{aligned}
L_{\mathcal{D}}(g) &= \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{1}_{g(x)\neq y}] = \mathbb{E}_{x\sim\mathcal{D}}\big[\mathbb{E}_{y\sim\mathcal{D}_{Y|x}}[\mathbb{1}_{g(x)\neq y}]\big] = \mathbb{E}_{x\sim\mathcal{D}}[\mathbb{P}_{y\sim\mathcal{D}_{Y|x}}(g(X) \neq Y | X = x)], \\
L_{\mathcal{D}}(f_{\mathcal{D}}) &= \mathbb{E}_{x\sim\mathcal{D}}[\mathbb{P}_{y\sim\mathcal{D}_{Y|x}}(f_{\mathcal{D}}(X) \neq Y | X = x)].
\end{aligned}
$$

We should compare the two conditional probabilities inside the expectation. Let $x \in \mathcal{X}$ and $a_x = \mathbb{P}(Y = 1|X = x)$. We have

$$\begin{aligned}
\mathbb{P}(g(X) \neq Y|X = x) &= \mathbb{P}(g(X) = 0|X = x) \cdot \mathbb{P}(Y = 1|X = x) \\
&+ \mathbb{P}(g(X) = 1|X = x) \cdot \mathbb{P}(Y = 0|X = x) \\
&= \mathbb{P}(g(X) = 0|X = x) \cdot a_x + \mathbb{P}(g(X) = 1|X = x) \cdot (1 - a_x) \\
&\geq \mathbb{P}(g(X) = 0|X = x) \cdot \min\{a_x, 1 - a_x\} \\
&+ \mathbb{P}(g(X) = 1|X = x) \cdot \min\{a_x, 1 - a_x\} \\
&= \min\{a_x, 1 - a_x\}.
\end{aligned}$$

When $g = f_{\mathcal{D}}$ we should replace $\mathbb{P}(g(X) = 0|X = x)$ by $\mathbb{1}_{a_x<1/2}$ and $\mathbb{P}(g(X) = 1|X = x)$ by $\mathbb{1}_{a_x\geq 1/2}$. Then the above inequality is tight:

$$\mathbb{P}(f_{\mathcal{D}}(X) \neq Y|X = x) = \mathbb{1}_{a_x<1/2} \cdot a_x + \mathbb{1}_{a_x\geq 1/2} \cdot (1 - a_x) = \min\{a_x, 1 - a_x\}.$$

Therefore, we have $L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$.

Exercise 3.8

1. Solved already in Exercise 3.7.
2. We have shown in Exercise 3.7 that the Bayes optimial predictor $f_{\mathcal{D}}$ is optimal w.r.t. $\mathcal{D}$; in other words, $f_{\mathcal{D}}$ is always better than any other learning algorithm w.r.t. $\mathcal{D}$.
3. Take $\mathcal{D}$ to be any probability distribution and $B = f_{\mathcal{D}}$.

Exercise 4.1

<u>$1 \Rightarrow 2$</u>: Assume for every $\epsilon, \delta > 0$ there exists $m(\epsilon, \delta)$ such that $\forall m \geq m(\epsilon, \delta)$

$$\mathbb{P}_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > \epsilon) < \delta. \tag{1}$$

Then using the definition of expectation

$$\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] &\leq \mathbb{P}_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > \epsilon) \cdot 1 + \mathbb{P}_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S)) \leq \epsilon) \cdot \epsilon \\
&\leq \mathbb{P}_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > \epsilon) + \epsilon \\
&\leq \delta + \epsilon,
\end{aligned}$$

where the last inequality follows from the assumption (1). Now set $\delta = \epsilon$. We have for every $\epsilon > 0$ there exists $m(\epsilon, \epsilon)$ such that $\forall m \geq m(\epsilon, \epsilon)$

$$\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \leq 2\epsilon. \tag{2}$$

So it is valid to pass both sides of (2) to the limit $\lim_{m\to\infty} \lim_{\epsilon\to 0}$, which gives

$$\lim_{m\to\infty} \mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \leq 0.$$

Also by definition $\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \geq 0$. Thus we conclude $\lim_{m\to\infty} \mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] = 0$.

$\underline{2 \Rightarrow 1}$: Assume that $\lim_{m\to\infty} \mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] = 0$. For every $\epsilon, \delta \in (0,1)$ there exists some $m_0 \in \mathbb{N}$ such that for every $m \geq m_0$, $\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \leq \epsilon\delta$. By Markov's inequality,

$$
\begin{aligned}
\mathbb{P}_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > \epsilon) &\leq \frac{\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))]}{\epsilon} \\
&\leq \frac{\epsilon\delta}{\epsilon} \\
&= \delta.
\end{aligned}
$$

Exercise 4.2

Using Hoeffding's inequality on $L_{\mathcal{D}} \in [a,b]$ we have

$$
\mathbb{P}_{S\sim\mathcal{D}^m}(|L_{\mathcal{D}}(h) - L_S(h)| > \epsilon) \leq 2\exp\Big(-\frac{2m\epsilon^2}{(b-a)^2}\Big).
$$

Then we substitute this into the step where the union bound is used:

$$
\begin{aligned}
\mathbb{P}_{S\sim\mathcal{D}^m}(\exists h \in \mathcal{H}, |L_S(h) - L_D(h)| > \epsilon) &\leq \sum_{h\in\mathcal{H}} \mathbb{P}_{S\sim\mathcal{D}^m}(|L_{\mathcal{D}}(h) - L_S(h)| > \epsilon) \\
&\leq 2|\mathcal{H}|\exp\Big(-\frac{2m\epsilon^2}{(b-a)^2}\Big)
\end{aligned}
$$

The desired bound on the sample complexity follows from requiring $2|\mathcal{H}|\exp\Big(-\frac{2m\epsilon^2}{(b-a)^2}\Big) \leq \delta$.