Problem #12.4: [The Halting Problem is Convex-Lipschitz-Bounded but Not Computable]

Suppose that $H = [0, 1]$ and $Z$ is the set of all Turing machines. Suppose the loss function is given by $\ell$

$$
\ell(0, T) = \begin{cases} 1, & \text{if } T \text{ halts} \\ 0, & \text{if } T \text{ does not halt.} \end{cases}
$$

$$
\ell(1, T) = \begin{cases} 0, & \text{if } T \text{ halts} \\ 1, & \text{if } T \text{ does not halt.} \end{cases}
$$

$$
\ell(h, T) = h\ell(0, T) + (1 - h)\ell(1, T), \text{ for } h \in H.
$$

Part (a): [Convex-Lipschitz-Bounded]

Show this learning problem is convex-Lipschitz-bounded.

**Solution**

No need to show that $H = [0, 1]$ is convex and bounded. For any $T \in Z$ the second derivative of $\ell(\cdot, T)$ is zero and hence convexity. Also $\ell(\cdot, T)$ is 1-Lipschitz because $\ell(h, T) = h\ell(0, T) + (1 - h)\ell(1, T)$ and for any $h_1, h_2 \in H, T \in Z$

$$
|\ell(h_1, T) - \ell(h_2, T)| = |(h_1 - h_2)(\ell(0, T) - \ell(1, T))| = |h_1 - h_2||\ell(0, T) - \ell(1, T)| = |h_1 - h_2|.
$$

Hence the conclusion.

Part (b): [Not Computable]

Show that no computable algorithm can learn this problem.

**Solution**

Suppose $A$ is an agnostic PAC learner of $H$, with sample complexity $m_H : (0, 1)^2 \to \mathbb{N}$. For any Turing machine $T$, denote $D_T$ as the distribution of the single Turing machine $T$. Then

$$
L_{D_T}(h) = \ell(h, T) = \begin{cases} h, & \text{if } T \text{ halts} \\ 1 - h, & \text{if } T \text{ does not halt.} \end{cases}
$$

So $\min_{h \in H} L_{D_T}(h) = 0$.

Fix $\epsilon \in (0, \frac{1}{2}), \delta \in (0, 1), T \in Z$. Now for $m \geq m_H(\epsilon, \delta), \mathbb{P}_{S \sim D_T}(\ell(A(S), T) \geq \epsilon) \leq \delta$ because $\min_{h \in H} L_{D_T}(h) = 0$. Without loss of generality $A(T) = A(\{T\}^m)$, so taking the limit as $\delta$ approaches 0 we conclude $\ell(A(S), T) \leq \epsilon$.

$$
\epsilon \geq \ell(A(S), T) = \begin{cases} A(S), & \text{if } T \text{ halts} \\ 1 - A(S), & \text{if } T \text{ does not halt.} \end{cases}
$$

Then $T$ halts if and only if $A(S) \leq \epsilon$. But no algorithm can determine if a general Turing machine halts. Thus $A$ is not computable.  
\[ \square \]
Problem #13.1: [From Bounded Expected Risk to Agnostic PAC Learning]

Let $A$ be an algorithm so that if $m \geq m_\epsilon(\epsilon)$, then for any $\mathcal{D}$ it holds that

$$\mathbb{E}_{S \sim \mathcal{D}}[L_D(A(S))] \leq \min_{h \in \mathcal{H}} L_D(h) + \epsilon.$$

(i) Show that for every $\delta > 0$, if $m \geq m_\epsilon(\epsilon \delta)$ then with probability at least $1 - \delta$ it holds that

$$L_D(A(S)) \leq \min_{h \in \mathcal{H}} L_D(h) + \epsilon.$$

(ii) For each $\delta \in (0,1)$ let $m_\delta(\epsilon, \delta) = m_\epsilon(\epsilon/2)(\lceil \log_2(1/\delta) \rceil + 1) + 2 \left[ \frac{\log(4/\delta) + \log(\log(1/\delta))}{\epsilon^2} \right]$. Suggest a procedure that agnostic PAC learns the problem with sample complexity $m_\delta(\epsilon, \delta)$, assuming the loss function is bounded by 1.

Solution

(i) Fix $\epsilon, \delta \in (0,1)$ and $m \geq m_\epsilon(\epsilon \delta)$. Using that $L_D(A(S)) - \min_{h \in \mathcal{H}} L_D(h) \geq 0$ and Markov’s inequality

$$\mathbb{P}_{S \sim \mathcal{D}}[\min_{h \in \mathcal{H}} L_D(h) \leq \epsilon] = 1 - \mathbb{P}_{S \sim \mathcal{D}}[\min_{h \in \mathcal{H}} L_D(h) > \epsilon] \geq 1 - \mathbb{E}_{S \sim \mathcal{D}}[\min_{h \in \mathcal{H}} L_D(h)]/\epsilon \geq 1 - \epsilon \delta/\epsilon = 1 - \delta.$$

(ii) Let $k = \lceil \log_2(1/\delta) \rceil + 1$. Fix $\epsilon, \delta(0,1)$ and let $m \geq m_\epsilon(\epsilon \delta)$. Let $S = \{S_1, \ldots, S_{k+1}\}$ where for each $i \leq k$ $|S_i| = m := m_\epsilon(\epsilon/2)$ and $|S_{k+1}| = m V := 2 \left[ \frac{\log(4/\delta) + \log(\log(1/\delta))}{\epsilon^2} \right]$. Denote $V = S_{k+1}$ for simplicity. Since $S_i$’s have distributions which are i.i.d. and by part (i) we have

$$\mathbb{P}_{S \sim \mathcal{D}}[L_D(A(S_i)) \geq \min_{h \in \mathcal{H}} L_D(h) + \epsilon \text{ for each } i] = \prod_{i=1}^{k} \mathbb{P}_{S_i \sim \mathcal{D}}[L_D(A(S_i)) \geq \min_{h \in \mathcal{H}} L_D(h) + \epsilon] \leq \prod_{i=1}^{k} \frac{1}{2} \leq 2^{-\log_2(1/\delta)-1} = \delta/2.$$

Consider the empirical hypothesis space $\mathcal{H}' = \{A(S_1), \ldots, A(S_k)\}$. Now

$$\mathbb{P}_{S \sim \mathcal{D}^{k+m} \mid V}[L_D(ERM_{\mathcal{H}'}(S_{k+1})) \leq \min_{h \in \mathcal{H}'} L_D(h) + \epsilon] = \mathbb{P}_{S \sim \mathcal{D}^{k+m} \mid V}[L_D(ERM_{\mathcal{H}'}(S_{k+1})) - \min_{h \in \mathcal{H}'} L_D(h) \leq \epsilon] \geq \mathbb{P}_{S \sim \mathcal{D}^{k+m} \mid V}[L_D(ERM_{\mathcal{H}'}(S_{k+1})) - \min_{h \in \mathcal{H}'} L_V(h) \leq \epsilon] \geq \mathbb{P}_{S \sim \mathcal{D}^{k+m} \mid V}[\min_{h \in \mathcal{H}'} L_V(h) \leq \epsilon] \geq \mathbb{P}_{V \sim \mathcal{D}^{m}}[\min_{h \in \mathcal{H}'} L_V(h) \leq \epsilon] = \mathbb{P}_{V \sim \mathcal{D}^{m}}[\min_{h \in \mathcal{H}'} L_D(h) \leq \epsilon].$$

Let $\delta' = 2 k / \epsilon^2 m V / 2$, then $\epsilon = 2 \sqrt{\frac{\log(2k/\delta')}{2mV}}$. Applying Theorem 11.2

$$\mathbb{P}_{S \sim \mathcal{D}^{k+m} \mid V}[L_D(ERM_{\mathcal{H}'}(S_{k+1})) \leq \min_{h \in \mathcal{H}'} L_D(h) + \epsilon] \geq (1 - \delta/2) \mathbb{P}_{V \sim \mathcal{D}^{m}}[L_D(h) - L_V(h) \leq 2 \sqrt{\frac{\log(2k/\delta')}{2mV}}, \text{ for all } h \in \mathcal{H}'] \geq (1 - \delta/2)(1 - \delta') \geq (1 - \delta/2)(1 - 2k / \epsilon^2 m V / 8) \geq (1 - \delta/2)(1 - 2 \log_2(1/\delta)/\epsilon^{\log(4/\delta) + \log(\log(1/\delta)))} = (1 - \delta/2)(1 - 1/2) > 1 - \delta.$$

$\square$
Remark 1. For problem #13.1 part 2, I was not able to prove as much as the book claims, and had an extra factor of 2 for the validation sets order, also the claim that $2^{-k} \leq \delta/2$ is false. So as you can see I gave a slightly larger sample complexity which I could handle. It may be due to my handling of the bound of $L_D \leq 1$. 
Problem #13.3: [Stability and Asymptotics ERM are Sufficient for Learnability]

Prove the following theorem.

**Theorem 2.** If a learning algorithm $A$ is on-average-replace-one-stable with rate $\epsilon_1(m)$ and is an AERM with rate $\epsilon_2(m)$, then is learns $\mathcal{H}$ with rate $\epsilon_1(m) + \epsilon_2(m)$.

**Solution**

By Theorem 13.2, $E_{S \sim D^m}[L_D(A(S)) - L_S(A(S))] \leq \epsilon_1(m)$. By definition of AERM

$E_{S \sim D^m}[L_S(A(S)) - \min_{h \in \mathcal{H}} L_S(h)] \leq \epsilon_2(m)$. So $E_{S \sim D^m}[L_D(A(S)) - \min_{h \in \mathcal{H}} L_S(h)] \leq \epsilon_1(m) + \epsilon_2(m)$.

By Jensen’s inequality $E_{S \sim D^m}[\min_{h \in \mathcal{H}} L_S(h)] \leq \min_{h \in \mathcal{H}} E_{S \sim D^m}[L_S(h)] = \min_{h \in \mathcal{H}} [L_D(h)]$. By total expectations $E_{S \sim D^m}[L_D(A(S)) - \min_{h \in \mathcal{H}} L_D(h)] \leq E_{S \sim D^m}[L_D(A(S)) - \min_{h \in \mathcal{H}} L_S(h)] \leq \epsilon_1(m) + \epsilon_2(m)$. □
Problem #14.1: [λ-Strongly Convex]

Prove claim 14.10. Hint: Extend proof of Lemma 13.5. If \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) is \( \lambda \)-strongly convex, then for all \( w, u \in \mathbb{R}^d \) and \( v \in \partial f(w) \), we have \( \langle w - u, v \rangle \geq f(w) - f(u) + \frac{\lambda}{2} \| w - u \|^2 \).

Solution

For any \( \alpha \in (0, 1) \),

\[
 f(\alpha u + (1 - \alpha)w) \leq \alpha f(u) + (1 - \alpha)f(w) - \frac{\lambda}{2} \alpha(1 - \alpha)\| w - u \|^2
\]

or

\[
 \frac{f(w) - f(\alpha u + (1 - \alpha)w)}{\alpha} \geq f(w) - f(u) + \frac{\lambda}{2} (1 - \alpha)\| w - u \|^2.
\]

Since \( v \in \partial f(w) \), \( f(\alpha u + (1 - \alpha)w) \geq f(w) + \langle \alpha u + (1 - \alpha)w - w, v \rangle = f(w) - \alpha \langle w - u, v \rangle \). This gives that \( \langle w - u, v \rangle \geq f(w) - f(u) + \frac{\lambda}{2} (1 - \alpha)\| w - u \|^2 \). Taking the limit as \( \alpha \rightarrow 0 \) gives the desired result. \( \square \)