Proof of the Fundamental Theorem

Peng Lingwei

August 6, 2019

Contents

28 Proof of the Fundamental Theorem of Learning Theory

28.1 THE UPPER BOUND FOR THE AGNOSTIC CASE

Nowadays, we have that $m_{\mathcal{H}}(\epsilon,\delta) \leq C\frac{d+\ln(1/\delta)}{\epsilon^2}$. But the proof need a careful analysis of the Rademacher complexity using a technique called "chaining".
In this chapter, we proof

$$m_{\mathcal{H}}(\epsilon,\delta) \leq C\frac{d\ln(d/\epsilon)+\ln(1/\delta)}{\epsilon^2}.$$

Proof. Let $\mathcal{H}_S = \{(h(\vec{x}_1),\ldots,h(\vec{x}_m)) : h \in \mathcal{H}, x_i \in S\}$, then $A = l^{0-1} \circ \mathcal{H}_S = \left\{(1_{y_1 \neq h(\vec{x}_1)},\ldots,1_{y_m \neq h(\vec{x}_m)}) : h \in \mathcal{H}, x_i \in S\right\}$.
By Sauer-Shelah lemma: $|A| = |\mathcal{H}_S| \leq \left(\frac{em}{d}\right)^d$.
By Massart lemma: $R(A) \leq \max_{\vec{a}\in A}\|\vec{a}-\bar{\vec{a}}\|\sqrt{2\ln(|A|)}/m = \sqrt{2\ln(|A|)/m}$.

$$\mathbb{P}\left\{|L_{\mathcal{D}}(h) - L_S(h)| \leq 2\mathbb{E}R(A) + \sqrt{2\ln(2/\delta)/m}\right\} \geq 1-\delta$$

$$\mathbb{P}\left\{|L_{\mathcal{D}}(h) - L_S(h)| \leq \sqrt{8d\ln(em/d)/m} + \sqrt{2\ln(2/\delta)/m}\right\} \geq 1-\delta$$

$$\mathbb{P}\left\{|L_{\mathcal{D}}(h) - L_S(h)| \leq \sqrt{16d\ln(em/d)/m + 4\ln(2/\delta)/m}\right\} \geq 1-\delta$$

Then we only need $m \geq \frac{16d}{\epsilon^2}\ln\left(\frac{em}{d}\right) + \frac{4}{\epsilon^2}\log(2/\delta)$.

$$m \geq \frac{16d}{\epsilon^2}\ln(m) + \frac{4}{\epsilon^2}\left(4d\ln(e/d) + \ln(2/\delta)\right)$$

we have that $\forall a > 0, b > 0, x \geq 4a\ln(2a) + 2b \Rightarrow x \geq a\ln(x) + b$. So, we only need

$$m \geq \frac{64d}{\epsilon^2}\ln\left(\frac{32d}{\epsilon^2}\right) + \frac{8}{\epsilon^2}\left(4d\ln(e/d) + \ln(2/\delta)\right)$$

Which means

$$m_{\mathcal{H}}(\epsilon,\delta) \leq \frac{64d}{\epsilon^2}\ln\left(\frac{32d}{\epsilon^2}\right) + \frac{8}{\epsilon^2}\left(4d\ln(e/d) + \ln(2/\delta)\right) \leq C\frac{d\ln(d/\epsilon)+\ln(1/\delta)}{\epsilon^2}$$

□

28.2 THE LOWER BOUND FOR THE AGNOSTIC CASE

This section's target is proofing $m_{\mathcal{H}}(\epsilon,\delta) \geq C\frac{d+\ln(1/\delta)}{\epsilon^2}$.

28.2.1 $m(\epsilon,\delta) \geq (1-\epsilon^2)/\epsilon^2 \log(1/(4\delta - 4\delta^2))$

$\mathcal{X} = \{c\}, \mathcal{Y} = \{+1,-1\}, \mathcal{H} = \{+1,-1\}, \mathbf{D} = \{\mathcal{D}_{+1},\mathcal{D}_{-1}\}$, where $\mathcal{D}_b = \frac{1+yb\epsilon}{2}$.
Let $S = \{(c,y_1),\ldots,(c,y_m)\}, \vec{y} = \{y_1,\ldots,y_m\}$.

$$\forall h \in \mathcal{H}, \quad L_{\mathcal{D}_b}(h) = \frac{1-h(c)b\epsilon}{2}.$$

So, the Bayes optimal hypothesis is $h_b(c) = b$.Then,

$$L_{\mathcal{D}_b}(A(\vec{y})) - \min_{h\in\mathcal{H}} L_{\mathcal{D}_b}(h_b) = \frac{1 - A(\vec{y})b\epsilon}{2} - \frac{1-\epsilon}{2} = \begin{cases} \epsilon & A(\vec{y}) \neq b \\ 0 & otherwise \end{cases}$$

$$\mathbb{P}_{\mathcal{D}_b}\left\{L_{\mathcal{D}_b}(A(\vec{y})) - \min_{h\in\mathcal{H}} L_{\mathcal{D}_b}(h_b) \geq \epsilon\right\} = \sum_{\vec{y}} \mathbb{P}_{\mathcal{D}_b}(\vec{y}) 1_{A(\vec{y})\neq b}$$

We denote $N^+ = \left\{\vec{y} : \langle \vec{1}, \vec{y}\rangle \geq 0\right\}$.

$$\begin{aligned}
&\max_{\mathcal{D}_b\in\mathbf{D}} \mathbb{P}_{\mathcal{D}_b}\left\{L_{\mathcal{D}_b}(A(\vec{y})) - \min_{h\in\mathcal{H}} L_{\mathcal{D}_b}(h_b) \geq \epsilon\right\} \\
&= \max_{\mathcal{D}_b\in\mathbf{D}} \sum_{\vec{y}} \mathbb{P}_{\mathcal{D}_b}[\vec{y}] 1_{[A(\vec{y})\neq b]} \\
&\geq \frac{1}{2}\sum_{\vec{y}} \mathbb{P}_{\mathcal{D}_{+1}}[\vec{y}] 1_{[A(\vec{y})\neq +1]} + \frac{1}{2}\sum_{\vec{y}} \mathbb{P}_{\mathcal{D}_{-1}}[\vec{y}] 1_{[A(\vec{y})\neq -1]} \\
&= \frac{1}{2}\sum_{\vec{y}\in N^+} \mathbb{P}_{\mathcal{D}_{+1}}[\vec{y}] 1_{[A(\vec{y})\neq +1]} + \frac{1}{2}\sum_{\vec{y}\in N^+} \mathbb{P}_{\mathcal{D}_{-1}}[\vec{y}] 1_{[A(\vec{y})\neq -1]} \\
&\quad \frac{1}{2}\sum_{\vec{y}\in N^-} \mathbb{P}_{\mathcal{D}_{+1}}[\vec{y}] 1_{[A(\vec{y})\neq +1]} + \frac{1}{2}\sum_{\vec{y}\in N^-} \mathbb{P}_{\mathcal{D}_{-1}}[\vec{y}] 1_{[A(\vec{y})\neq -1]} \\
&\geq \frac{1}{2}\sum_{\vec{y}\in N^+} \mathbb{P}_{\mathcal{D}_{-1}}[\vec{y}] 1_{[A(\vec{y})\neq +1]} + \frac{1}{2}\sum_{\vec{y}\in N^+} \mathbb{P}_{\mathcal{D}_{-1}}[\vec{y}] 1_{[A(\vec{y})\neq -1]} \\
&\quad \frac{1}{2}\sum_{\vec{y}\in N^-} \mathbb{P}_{\mathcal{D}_{+1}}[\vec{y}] 1_{[A(\vec{y})\neq +1]} + \frac{1}{2}\sum_{\vec{y}\in N^-} \mathbb{P}_{\mathcal{D}_{+1}}[\vec{y}] 1_{[A(\vec{y})\neq -1]} \\
&= \frac{1}{2}\sum_{\vec{y}\in N^+} P_{\mathcal{D}_{-1}}[\vec{y}] + \frac{1}{2}\sum_{\vec{y}\in N^-} P_{\mathcal{D}_{+1}}[\vec{y}] = \sum_{\vec{y}\in N^-} P_{\mathcal{D}_{+1}}[\vec{y}]
\end{aligned}$$

The probability equals the probability that a Binomial $(m, (1-\epsilon)/2)$ random variable will have value greater than m/2. Using Slud's inequality, we have

$$\sum_{\vec{y}\in N^-} p_{\mathcal{D}_{+1}}[\vec{y}] \geq \frac{1}{2}\left(1 - \sqrt{1 - \exp\left(-m\epsilon^2/(1-\epsilon^2)\right)}\right) \geq \delta$$

$$m \leq \frac{1-\epsilon^2}{\epsilon^2} \ln\frac{1}{4\delta - 4\delta^2} \Rightarrow m_{\mathcal{H}}(\epsilon,\delta) \geq \frac{1-\epsilon^2}{\epsilon^2} \ln\frac{1}{4\delta - 4\delta^2} \geq C\frac{\ln(1/\delta)}{\epsilon^2}$$

28.2.2 Showing That $m(\epsilon, \delta) \geq d/(32\epsilon^2)$

Let $\mathcal{X} = \{x_1, \ldots, x_d\}$, $\mathcal{Y} = \{+1, -1\}$, and $\mathcal{H}$ shatters $\mathcal{X}$.
We only consider $\mathbf{D}_\rho = \left\{\mathcal{D}_{\vec{b}} : \vec{b} \in \{\pm 1\}^d\right\}$, where

$$\mathcal{D}_{\vec{b}}(\{(x,y)\}) \begin{cases} \frac{1}{d}\cdot\frac{1+yb_i\rho}{2} & \exists i : x = c_i \\ 0 & otherwise. \end{cases}$$

$$\forall h \in \mathcal{H}, L_{\mathcal{D}_{\vec{b}}}(h) = \frac{1+\rho}{2} \cdot \frac{|\{i \in [d] : h(c_i) \neq b_i\}|}{d} + \frac{1-\rho}{2} \cdot \frac{|\{i \in [d] : h(c_i) = b_i\}|}{d}$$

$$\min_{h \in \mathcal{H}} L_{\mathcal{D}_{\vec{b}}}(h) = \frac{1-\rho}{2} \Rightarrow L_{\mathcal{D}_{\vec{b}}}(h) - \min_{h \in \mathcal{H}} L_{\mathcal{D}_{\vec{b}}}(h) = \rho \cdot \frac{|\{i \in [d] : h(c_i) \neq b_i\}|}{d}.$$

which means that

$$L_{\mathcal{D}_{\vec{b}}}(h) - \min_{h \in \mathcal{H}} L_{\mathcal{D}_{\vec{b}}}(h) \in [0, \rho]$$

$$\begin{aligned}
&\max_{\mathcal{D}_{\vec{b}} \in \mathbf{D}_\rho} \mathbb{E}_{S \sim \mathcal{D}_{\vec{b}}^m} \left[L_{\mathcal{D}_{\vec{b}}}(A(S)) - \min_{h \in \mathcal{H}} L_{\mathcal{D}_{\vec{b}}}(h) \right] \\
\geq & \mathbb{E}_{\mathcal{D}_{\vec{b}} \sim U(\mathbf{D}_\rho)} \mathbb{E}_{S \sim \mathcal{D}_{\vec{b}}^m} \left[L_{\mathcal{D}_{\vec{b}}}(A(S)) - \min_{h \in \mathcal{H}} L_{\mathcal{D}_{\vec{b}}}(h) \right] \\
= & \mathbb{E}_{\mathcal{D}_{\vec{b}} \sim U(\mathbf{D}_\rho)} \mathbb{E}_{S \sim \mathcal{D}_{\vec{b}}^m} \left[\rho \cdot \frac{|\{i \in [d] : A(S)(c_i) \neq b_i\}|}{d} \right] \\
= & \frac{\rho}{d} \sum_{i=1}^{d} \mathbb{E}_{\mathcal{D}_{\vec{b}} \sim U(\mathbf{D}_\rho)} \mathbb{E}_{S \sim \mathcal{D}_{\vec{b}}^m} 1_{[A(S)(c_i) \neq b_i]} \\
= & \frac{\rho}{d} \sum_{i=1}^{d} \mathbb{E}_{\vec{j} \sim U([d])^m} \mathbb{E}_{\vec{b} \sim \{\pm 1\}^m} \mathbb{E}_{\vec{y} \sim b_{\vec{j}}} 1_{[A(c_{\vec{j}}, \vec{y})(c_i) \neq b_i]} \\
& \mathbb{E}_{\vec{b} \sim \{\pm 1\}^m} \mathbb{E}_{\vec{y} \sim b_{\vec{j}}} 1_{[A(c_{\vec{j}}, \vec{y})(c_i) \neq b_i]} \\
= & \mathbb{E}_{(\vec{b} - b_i) \sim \{\pm 1\}^{m-1}} \mathbb{E}_{\vec{y}^{\neg I} \sim b_{\vec{j}}^{\neg I}} \mathbb{E}_{b_i \sim \{\pm 1\}} \mathbb{E}_{\vec{y}^I \sim b_i} 1_{[A(c_{\vec{j}}, \vec{y})(c_i) \neq b_i]} \\
= & \mathbb{E}_{(\vec{b} - b_i) \sim \{\pm 1\}^{m-1}} \mathbb{E}_{\vec{y}^{\neg I} \sim b_{\vec{j}}^{\neg I}} \left[\frac{1}{2} \sum_{y^I} \left(\sum_{b_i \in \{\pm 1\}} \mathbb{P}[y^I | b_i] 1_{[A(c_{\vec{j}}, \vec{y})(c_i) \neq b_i]} \right) \right] \\
\geq & \mathbb{E}_{(\vec{b} - b_i) \sim \{\pm 1\}^{m-1}} \mathbb{E}_{\vec{y}^{\neg I} \sim b_{\vec{j}}^{\neg I}} \left[\frac{1}{2} \sum_{y^I} \left(\sum_{b_i \in \{\pm 1\}} \mathbb{P}[y^I | b_i] 1_{[A_{ML}(c_{\vec{j}}, \vec{y})(c_i) \neq b_i]} \right) \right]
\end{aligned}$$

where $A_{ML}(S)(c_i) = sign\left(\sum_{r : x_r = c_i} y_r\right)$. In equation

$$\mathbb{E}_{\vec{b} \sim \{\pm 1\}^m} \mathbb{E}_{\vec{y} \sim b_{\vec{j}}} 1_{[A(c_{\vec{j}}, \vec{y})(c_i) \neq b_i]}$$

we fix the $X = \{x_1, \ldots, x_m\}$'s index vector $\vec{j}$.We denote $n_{\vec{j}}(i)$ as the number i occurring in $\vec{j}$. We want maximum-likelihood going wrong, which means that $B \sim (n_{\vec{j}}(i), (1-\rho)/2) \geq n_{\vec{j}}(i)/2$ occuring.

$$\mathbb{P}\left[B \geq n_{\vec{j}}(i)/2\right] \geq \frac{1}{2} \left(1 - \sqrt{1 - \exp\left\{-2 n_{\vec{j}}(i) \rho^2\right\}}\right)$$

$$
\begin{aligned}
&\max_{\mathcal{D}_{\vec{b}} \in \mathbf{D}_\rho} \mathbb{E}_{S \sim \mathcal{D}_{\vec{b}}^m} \left[L_{\mathcal{D}_{\vec{b}}}(A(S)) - \min_{h \in \mathcal{H}} L_{\mathcal{D}_{\vec{b}}}(h) \right] \\
\geq & \frac{\rho}{2d} \sum_{i=1}^{d} \mathbb{E}_{\vec{j} \sim U([d])^m} \left(1 - \sqrt{1 - \exp\left\{-2 n_{\vec{j}}(i) \rho^2\right\}} \right) \\
\geq & \frac{\rho}{2d} \sum_{i=1}^{d} \left(1 - \sqrt{1 - \exp\left\{-2\rho^2 \mathbb{E}_{\vec{j} \sim U([d])^m} n_{\vec{j}}(i)\right\}} \right) \\
= & \frac{\rho}{2d} \sum_{i=1}^{d} \left(1 - \sqrt{1 - \exp\left\{-2\rho^2 m/d\right\}} \right) \\
= & \frac{\rho}{2} \left(1 - \sqrt{1 - \exp\left\{-2\rho^2 m/d\right\}} \right) \\
\geq & \frac{\rho}{2} \left(1 - \sqrt{2\rho^2 m/d} \right) \\
&\max_{\rho} \max_{\mathcal{D} \in \mathbf{D}_\rho} \mathbb{P}_{\mathcal{D}} \left[L_{\mathcal{D}}(A(S)) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) > \epsilon \right] \\
= & \max_{\rho} \max_{\mathcal{D} \in \mathbf{D}_\rho} \mathbb{P}_{\mathcal{D} \in \mathbf{D}_\rho} \left[\frac{1}{\rho} \left(L_{\mathcal{D}}(A(S)) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \right) > \frac{\epsilon}{\rho} \right] \\
\geq & \max_{\rho} \max_{\mathcal{D} \in \mathbf{D}_\rho} \mathbb{E} \left[\frac{1}{\rho} \left(L_{\mathcal{D}}(A(S)) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \right) \right] - \frac{\epsilon}{\rho} \\
\geq & \max_{\rho} \frac{1}{2} \left(1 - \sqrt{2\rho^2 m/d} \right) - \frac{\epsilon}{\rho} = \max_{\rho} \frac{1}{2} - \left(\rho \sqrt{\frac{m}{2d}} + \frac{\epsilon}{\rho} \right) \\
= & \frac{1}{2} - 2\sqrt{\epsilon \sqrt{m/(2d)}} \geq \delta \Rightarrow m \leq \frac{d(1-2\delta)^2}{8\epsilon^2}
\end{aligned}
$$

Overall, $m_{\mathcal{H}}(\epsilon, \delta) \geq \frac{d(1-2\delta)^2}{8\epsilon^2}$. In reality, we want δ as small as possible, we can constrain $\delta \in (0, 1/4)$, then $m_{\mathcal{H}}(\epsilon, \delta) \geq \frac{d}{32\epsilon^2}$.

28.3 THE UPPER BOUND FOR THE REALIZABLE CASE

The sample complexity of PAC learnable:

$$
m_{\mathcal{H}}(\epsilon, \delta) \leq C \frac{d \ln(1/\epsilon) + \ln(1/\delta)}{\epsilon}.
$$