28

Proof of the Fundamental Theorem of Learning Theory

In this chapter we prove Theorem 6.8 from Chapter 6. We remind the reader the conditions of the theorem, which will hold throughout this chapter: $\mathcal{H}$ is a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0,1\}$, the loss function is the $0-1$ loss, and $\text{VCdim}(\mathcal{H}) = d < \infty$.

We shall prove the upper bound for both the realizable and agnostic cases and shall prove the lower bound for the agnostic case. The lower bound for the realizable case is left as an exercise.

28.1 THE UPPER BOUND FOR THE AGNOSTIC CASE

For the upper bound we need to prove that there exists C such that $\mathcal{H}$ is agnostic PAC learnable with sample complexity

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

We will prove the slightly looser bound:

$$m_{\mathcal{H}}(\epsilon, \delta) \leq C \frac{d \log(d/\epsilon) + \ln(1/\delta)}{\epsilon^2}. \tag{28.1}$$

The tighter bound in the theorem statement requires a more involved proof, in which a more careful analysis of the Rademacher complexity using a technique called "chaining" should be used. This is beyond the scope of this book.

To prove Equation (28.1), it suffices to show that applying the ERM with a sample size

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

yields an ϵ, δ-learner for $\mathcal{H}$. We prove this result on the basis of Theorem 26.5.

Let $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ be a classification training set. Recall that the Sauer-Shelah lemma tells us that if $\text{VCdim}(\mathcal{H}) = d$ then

$$\left|\{(h(\mathbf{x}_1), \ldots, h(\mathbf{x}_m)) : h \in \mathcal{H}\}\right| \leq \left(\frac{e\,m}{d}\right)^d.$$

Denote $A = \{(\mathbb{1}_{[h(\mathbf{x}_1)\neq y_1]}, \ldots, \mathbb{1}_{[h(\mathbf{x}_m)\neq y_m]}) : h \in \mathcal{H}\}$. This clearly implies that

$$|A| \leq \left(\frac{em}{d}\right)^d.$$

Combining this with Lemma 26.8 we obtain the following bound on the Rademacher complexity:

$$R(A) \leq \sqrt{\frac{2d\log(em/d)}{m}}.$$

Using Theorem 26.5 we obtain that with probability of at least $1-\delta$, for every $h \in \mathcal{H}$ we have that

$$L_{\mathcal{D}}(h) - L_S(h) \leq \sqrt{\frac{8d\log(em/d)}{m}} + \sqrt{\frac{2\log(2/\delta)}{m}}.$$

Repeating the previous argument for minus the zero-one loss and applying the union bound we obtain that with probability of at least $1-\delta$, for every $h \in \mathcal{H}$ it holds that

$$\begin{aligned}|L_{\mathcal{D}}(h) - L_S(h)| &\leq \sqrt{\frac{8d\log(em/d)}{m}} + \sqrt{\frac{2\log(4/\delta)}{m}} \\ &\leq 2\sqrt{\frac{8d\log(em/d) + 2\log(4/\delta)}{m}}.\end{aligned}$$

To ensure that this is smaller than ϵ we need

$$m \geq \frac{4}{\epsilon^2} \cdot \left(8d\log(m) + 8d\log(e/d) + 2\log(4/\delta)\right).$$

Using Lemma A.2, a sufficient condition for the inequality to hold is that

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

28.2 THE LOWER BOUND FOR THE AGNOSTIC CASE

Here, we prove that there exists C such that $\mathcal{H}$ is agnostic PAC learnable with sample complexity

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

We will prove the lower bound in two parts. First, we will show that $m(\epsilon, \delta) \geq 0.5\log(1/(4\delta))/\epsilon^2$, and second we will show that for every $\delta \leq 1/8$ we have that $m(\epsilon, \delta) \geq 8d/\epsilon^2$. These two bounds will conclude the proof.

28.2.1 Showing That $m(\epsilon, \delta) \geq 0.5\log(1/(4\delta))/\epsilon^2$

We first show that for any $\epsilon < 1/\sqrt{2}$ and any $\delta \in (0,1)$, we have that $m(\epsilon, \delta) \geq 0.5\log(1/(4\delta))/\epsilon^2$. To do so, we show that for $m \leq 0.5\log(1/(4\delta))/\epsilon^2$, $\mathcal{H}$ is not learnable.

Choose one example that is shattered by $\mathcal{H}$. That is, let c be an example such that there are $h_+, h_- \in \mathcal{H}$ for which $h_+(c) = 1$ and $h_-(c) = -1$. Define two distributions,

$\mathcal{D}_+$ and $\mathcal{D}_-$, such that for $b \in \{\pm 1\}$ we have

$$\mathcal{D}_b(\{(x,y)\}) = \begin{cases} \frac{1+yb\epsilon}{2} & \text{if } x = c \\ 0 & \text{otherwise.} \end{cases}$$

That is, all the distribution mass is concentrated on two examples $(c,1)$ and $(c,-1)$, where the probability of (c,b) is $\frac{1+b\epsilon}{2}$ and the probability of $(c,-b)$ is $\frac{1-b\epsilon}{2}$.

Let A be an arbitrary algorithm. Any training set sampled from $\mathcal{D}_b$ has the form $S = (c,y_1),\ldots,(c,y_m)$. Therefore, it is fully characterized by the vector $\mathbf{y} = (y_1,\ldots,y_m) \in \{\pm 1\}^m$. Upon receiving a training set S, the algorithm A returns a hypothesis $h : \mathcal{X} \to \{\pm 1\}$. Since the error of A w.r.t. $\mathcal{D}_b$ only depends on $h(c)$, we can think of A as a mapping from $\{\pm 1\}^m$ into $\{\pm 1\}$. Therefore, we denote by $A(\mathbf{y})$ the value in $\{\pm 1\}$ corresponding to the prediction of $h(c)$, where h is the hypothesis that A outputs upon receiving the training set $S = (c,y_1),\ldots,(c,y_m)$.

Note that for any hypothesis h we have

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

In particular, the Bayes optimal hypothesis is h_b and

$$L_{\mathcal{D}_b}(A(\mathbf{y})) - L_{\mathcal{D}_b}(h_b) = \frac{1 - A(\mathbf{y})b\epsilon}{2} - \frac{1-\epsilon}{2} = \begin{cases} \epsilon & \text{if } A(\mathbf{y}) \neq b \\ 0 & \text{otherwise.} \end{cases}$$

Fix A. For $b \in \{\pm 1\}$, let $Y^b = \{\mathbf{y} \in \{0,1\}^m : A(\mathbf{y}) \neq b\}$. The distribution $\mathcal{D}_b$ induces a probability P_b over $\{\pm 1\}^m$. Hence,

$$\mathbb{P}[L_{\mathcal{D}_b}(A(\mathbf{y})) - L_{\mathcal{D}_b}(h_b) = \epsilon] = \mathcal{D}_b(Y^b) = \sum_{\mathbf{y}} P_b[\mathbf{y}] \mathbb{1}_{[A(\mathbf{y}) \neq b]}.$$

Denote $N^+ = \{\mathbf{y} : |\{i : y_i = 1\}| \geq m/2\}$ and $N^- = \{\pm 1\}^m \setminus N^+$. Note that for any $\mathbf{y} \in N^+$ we have $P_+[\mathbf{y}] \geq P_-[\mathbf{y}]$ and for any $\mathbf{y} \in N^-$ we have $P_-[\mathbf{y}] \geq P_+[\mathbf{y}]$. Therefore,

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

$$\geq \frac{1}{2}\sum_{\mathbf{y}\in N^+}\left(P_-[\mathbf{y}]\mathbb{1}_{[A(\mathbf{y})\neq+]}+P_-[\mathbf{y}]\mathbb{1}_{[A(\mathbf{y})\neq-]}\right)$$
$$+\frac{1}{2}\sum_{\mathbf{y}\in N^-}\left(P_+[\mathbf{y}]\mathbb{1}_{[A(\mathbf{y})\neq+]}+P_+[\mathbf{y}]\mathbb{1}_{[A(\mathbf{y})\neq-]}\right)$$
$$=\frac{1}{2}\sum_{\mathbf{y}\in N^+}P_-[\mathbf{y}]+\frac{1}{2}\sum_{\mathbf{y}\in N^-}P_+[\mathbf{y}].$$

Next note that $\sum_{\mathbf{y}\in N^+}P_-[\mathbf{y}]=\sum_{\mathbf{y}\in N^-}P_+[\mathbf{y}]$, and both values are the probability that a Binomial $(m,(1-\epsilon)/2)$ random variable will have value greater than $m/2$. Using Lemma B.11, this probability is lower bounded by

$$\frac{1}{2}\left(1-\sqrt{1-\exp(-m\epsilon^2/(1-\epsilon^2))}\right)\ \geq\ \frac{1}{2}\left(1-\sqrt{1-\exp(-2m\epsilon^2)}\right),$$

where we used the assumption that $\epsilon^2\leq 1/2$. It follows that if $m\leq 0.5\log(1/(4\delta))/\epsilon^2$ then there exists b such that

$$\mathbb{P}[L_{\mathcal{D}_b}(A(\mathbf{y}))-L_{\mathcal{D}_b}(h_b)=\epsilon]$$
$$\geq\frac{1}{2}\left(1-\sqrt{1-\sqrt{4\delta}}\right)\geq\delta,$$

where the last inequality follows by standard algebraic manipulations. This concludes our proof.

28.2.2 Showing That $m(\epsilon,1/8)\geq 8d/\epsilon^2$

We shall now prove that for every $\epsilon<1/(8\sqrt{2})$ we have that $m(\epsilon,\delta)\geq\frac{8d}{\epsilon^2}$.

Let $\rho=8\epsilon$ and note that $\rho\in(0,1/\sqrt{2})$. We will construct a family of distributions as follows. First, let $C=\{c_1,\ldots,c_d\}$ be a set of d instances which are shattered by $\mathcal{H}$. Second, for each vector $(b_1,\ldots,b_d)\in\{\pm1\}^d$, define a distribution $\mathcal{D}_b$ such that

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

That is, to sample an example according to $\mathcal{D}_b$, we first sample an element $c_i\in C$ uniformly at random, and then set the label to be b_i with probability $(1+\rho)/2$ or $-b_i$ with probability $(1-\rho)/2$.

It is easy to verify that the Bayes optimal predictor for $\mathcal{D}_b$ is the hypothesis $h\in\mathcal{H}$ such that $h(c_i)=b_i$ for all $i\in[d]$, and its error is $\frac{1-\rho}{2}$. In addition, for any other function $f:\mathcal{X}\to\{\pm1\}$, it is easy to verify that

$$L_{\mathcal{D}_b}(f)=\frac{1+\rho}{2}\cdot\frac{|\{i\in[d]:f(c_i)\neq b_i\}|}{d}+\frac{1-\rho}{2}\cdot\frac{|\{i\in[d]:f(c_i)=b_i\}|}{d}.$$

Therefore,

$$L_{\mathcal{D}_b}(f)-\min_{h\in\mathcal{H}}L_{\mathcal{D}_b}(h)=\rho\cdot\frac{|\{i\in[d]:f(c_i)\neq b_i\}|}{d}. \tag{28.2}$$

Next, fix some learning algorithm A. As in the proof of the No-Free-Lunch theorem, we have that

$$\max_{\mathcal{D}_b : b \in \{\pm 1\}^d} \mathop{\mathbb{E}}_{S \sim \mathcal{D}_b^m} \left[L_{\mathcal{D}_b}(A(S)) - \min_{h \in \mathcal{H}} L_{\mathcal{D}_b}(h) \right] \tag{28.3}$$

$$\geq \mathop{\mathbb{E}}_{\mathcal{D}_b : b \sim U(\{\pm 1\}^d)} \mathop{\mathbb{E}}_{S \sim \mathcal{D}_b^m} \left[L_{\mathcal{D}_b}(A(S)) - \min_{h \in \mathcal{H}} L_{\mathcal{D}_b}(h) \right] \tag{28.4}$$

$$= \mathop{\mathbb{E}}_{\mathcal{D}_b : b \sim U(\{\pm 1\}^d)} \mathop{\mathbb{E}}_{S \sim \mathcal{D}_b^m} \left[\rho \cdot \frac{|\{i \in [d] : A(S)(c_i) \neq b_i|}{d} \right] \tag{28.5}$$

$$= \frac{\rho}{d} \sum_{i=1}^{d} \mathop{\mathbb{E}}_{\mathcal{D}_b : b \sim U(\{\pm 1\}^d)} \mathop{\mathbb{E}}_{S \sim \mathcal{D}_b^m} \mathbb{1}_{[A(S)(c_i) \neq b_i]}, \tag{28.6}$$

where the first equality follows from Equation (28.2). In addition, using the definition of $\mathcal{D}_b$, to sample $S \sim \mathcal{D}_b$ we can first sample $(j_1, \ldots, j_m) \sim U([d])^m$, set $x_r = c_{j_i}$, and finally sample y_r such that $\mathbb{P}[y_r = b_{j_i}] = (1+\rho)/2$. Let us simplify the notation and use $y \sim b$ to denote sampling according to $\mathbb{P}[y = b] = (1+\rho)/2$. Therefore, the right-hand side of Equation (28.6) equals

$$\frac{\rho}{d} \sum_{i=1}^{d} \mathop{\mathbb{E}}_{j \sim U([d])^m} \mathop{\mathbb{E}}_{b \sim U(\{\pm 1\}^d)} \mathop{\mathbb{E}}_{\forall r, y_r \sim b_{j_r}} \mathbb{1}_{[A(S)(c_i) \neq b_i]}. \tag{28.7}$$

We now proceed in two steps. First, we show that among all learning algorithms, A, the one which minimizes Equation (28.7) (and hence also Equation (28.4)) is the Maximum-Likelihood learning rule, denoted A_{ML}. Formally, for each i, $A_{ML}(S)(c_i)$ is the majority vote among the set $\{y_r : r \in [m], x_r = c_i\}$. Second, we lower bound Equation (28.7) for A_{ML}.

Lemma 28.1. *Among all algorithms, Equation (28.4) is minimized for A being the Maximum-Likelihood algorithm, A_{ML}, defined as*

$$\forall i, \quad A_{ML}(S)(c_i) = sign\left(\sum_{r : x_r = c_i} y_r \right).$$

Proof. Fix some $j \in [d]^m$. Note that given j and $y \in \{\pm 1\}^m$, the training set S is fully determined. Therefore, we can write $A(j, y)$ instead of $A(S)$. Let us also fix $i \in [d]$. Denote $b^{\neg i}$ the sequence $(b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_m)$. Also, for any $y \in \{\pm 1\}^m$, let y^I denote the elements of y corresponding to indices for which $j_r = i$ and let $y^{\neg I}$ be the rest of the elements of y. We have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{b \sim U(\{\pm 1\}^d)} \mathop{\mathbb{E}}_{\forall r, y_r \sim b_{j_r}} \mathbb{1}_{[A(S)(c_i) \neq b_i]} \\
&= \frac{1}{2} \sum_{b_i \in \{\pm 1\}} \mathop{\mathbb{E}}_{b^{\neg i} \sim U(\{\pm 1\}^{d-1})} \sum_{y} P[y | b^{\neg i}, b_i] \mathbb{1}_{[A(j,y)(c_i) \neq b_i]} \\
&= \mathop{\mathbb{E}}_{b^{\neg i} \sim U(\{\pm 1\}^{d-1})} \sum_{y^{\neg I}} P[y^{\neg I} | b^{\neg i}] \frac{1}{2} \sum_{y^I} \left(\sum_{b_i \in \{\pm 1\}} P[y^I | b_i] \mathbb{1}_{[A(j,y)(c_i) \neq b_i]} \right).
\end{aligned}$$

The sum within the parentheses is minimized when $A(j, y)(c_i)$ is the maximizer of $P[y^I|b_i]$ over $b_i \in \{\pm 1\}$, which is exactly the Maximum-Likelihood rule. Repeating the same argument for all i we conclude our proof. □

Fix i. For every j, let $n_i(j) = \{|t : j_t = i|\}$ be the number of instances in which the instance is c_i. For the Maximum-Likelihood rule, we have that the quantity

$$\mathop{\mathbb{E}}_{b\sim U(\{\pm 1\}^d)} \mathop{\mathbb{E}}_{\forall r, y_r \sim b_{j_r}} \mathbb{1}_{[A_{ML}(S)(c_i)\neq b_i]}$$

is exactly the probability that a binomial $(n_i(j), (1-\rho)/2)$ random variable will be larger than $n_i(j)/2$. Using Lemma B.11, and the assumption $\rho^2 \leq 1/2$, we have that

$$P[B \geq n_i(j)/2] \geq \frac{1}{2}\left(1 - \sqrt{1 - e^{-2n_i(j)\rho^2}}\right).$$

We have thus shown that

$$\begin{aligned}
&\frac{\rho}{d}\sum_{i=1}^{d} \mathop{\mathbb{E}}_{j\sim U([d])^m} \mathop{\mathbb{E}}_{b\sim U(\{\pm 1\}^d)} \mathop{\mathbb{E}}_{\forall r, y_r \sim b_{j_r}} \mathbb{1}_{[A(S)(c_i)\neq b_i]} \\
&\quad \geq \frac{\rho}{2d}\sum_{i=1}^{d} \mathop{\mathbb{E}}_{j\sim U([d])^m} \left(1 - \sqrt{1 - e^{-2\rho^2 n_i(j)}}\right) \\
&\quad \geq \frac{\rho}{2d}\sum_{i=1}^{d} \mathop{\mathbb{E}}_{j\sim U([d])^m} \left(1 - \sqrt{2\rho^2 n_i(j)}\right),
\end{aligned}$$

where in the last inequality we used the inequality $1 - e^{-a} \leq a$.

Since the square root function is concave, we can apply Jensen's inequality to obtain that the above is lower bounded by

$$\begin{aligned}
&\geq \frac{\rho}{2d}\sum_{i=1}^{d} \left(1 - \sqrt{2\rho^2 \mathop{\mathbb{E}}_{j\sim U([d])^m} n_i(j)}\right) \\
&= \frac{\rho}{2d}\sum_{i=1}^{d} \left(1 - \sqrt{2\rho^2 m/d}\right) \\
&= \frac{\rho}{2}\left(1 - \sqrt{2\rho^2 m/d}\right).
\end{aligned}$$

As long as $m < \frac{d}{8\rho^2}$, this term would be larger than $\rho/4$.

In summary, we have shown that if $m < \frac{d}{8\rho^2}$ then for any algorithm there exists a distribution such that

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(A(S)) - \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h)\right] \geq \rho/4.$$

Finally, Let $\Delta = \frac{1}{\rho}(L_{\mathcal{D}}(A(S)) - \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h))$ and note that $\Delta \in [0,1]$ (see Equation (28.5)). Therefore, using Lemma B.1, we get that

$$\mathbb{P}[L_{\mathcal{D}}(A(S)) - \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) > \epsilon] = \mathbb{P}\left[\Delta > \frac{\epsilon}{\rho}\right] \geq \mathbb{E}[\Delta] - \frac{\epsilon}{\rho}$$
$$\geq \frac{1}{4} - \frac{\epsilon}{\rho}.$$

Choosing $\rho = 8\epsilon$ we conclude that if $m < \frac{8d}{\epsilon^2}$, then with probability of at least $1/8$ we will have $L_{\mathcal{D}}(A(S)) - \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) \geq \epsilon$.

28.3 THE UPPER BOUND FOR THE REALIZABLE CASE

Here we prove that there exists C such that $\mathcal{H}$ is PAC learnable with sample complexity

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

We do so by showing that for $m \geq C\frac{d\ln(1/\epsilon)+\ln(1/\delta)}{\epsilon}$, $\mathcal{H}$ is learnable using the ERM rule. We prove this claim on the basis of the notion of ϵ-nets.

Definition 28.2 (ϵ-net). Let $\mathcal{X}$ be a domain. $S \subset \mathcal{X}$ is an ϵ-net for $\mathcal{H} \subset 2^{\mathcal{X}}$ with respect to a distribution $\mathcal{D}$ over $\mathcal{X}$ if

$$\forall h \in \mathcal{H}: \ \mathcal{D}(h) \geq \epsilon \ \Rightarrow \ h \cap S \neq \emptyset.$$

Theorem 28.3. *Let $\mathcal{H} \subset 2^{\mathcal{X}}$ with $\mathrm{VCdim}(\mathcal{H}) = d$. Fix $\epsilon \in (0,1)$, $\delta \in (0,1/4)$ and let*

$$m \geq \frac{8}{\epsilon}\left(2d\log\left(\frac{16e}{\epsilon}\right) + \log\left(\frac{2}{\delta}\right)\right).$$

Then, with probability of at least $1-\delta$ over a choice of $S \sim \mathcal{D}^m$ we have that S is an ϵ-net for $\mathcal{H}$.

Proof. Let

$$B = \{S \subset \mathcal{X} : |S| = m, \ \exists h \in \mathcal{H}, \mathcal{D}(h) \geq \epsilon, h \cap S = \emptyset\}$$

be the set of sets which are not ϵ-nets. We need to bound $\mathbb{P}[S \in B]$. Define

$$B' = \{(S,T) \subset \mathcal{X} : |S| = |T| = m, \ \exists h \in \mathcal{H}, \mathcal{D}(h) \geq \epsilon, h \cap S = \emptyset, \ |T \cap h| > \tfrac{\epsilon m}{2}\}.$$

Claim 1
$\mathbb{P}[S \in B] \leq 2\mathbb{P}[(S,T) \in B']$.
Proof of Claim 1: Since S and T are chosen independently we can write

$$\mathbb{P}[(S,T) \in B'] = \mathop{\mathbb{E}}_{(S,T)\sim\mathcal{D}^{2m}}\left[\mathbb{1}_{[(S,T)\in B']}\right] = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\mathop{\mathbb{E}}_{T\sim\mathcal{D}^m}\left[\mathbb{1}_{[(S,T)\in B']}\right]\right].$$

Note that $(S,T) \in B'$ implies $S \in B$ and therefore $\mathbb{1}_{[(S,T)\in B']} = \mathbb{1}_{[(S,T)\in B']}\mathbb{1}_{[S\in B]}$, which gives

$$\begin{aligned}\mathbb{P}[(S,T)\in B'] &= \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{T\sim\mathcal{D}^m}\mathbb{1}_{[(S,T)\in B']}\mathbb{1}_{[S\in B]} \\ &= \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathbb{1}_{[S\in B]}\mathop{\mathbb{E}}_{T\sim\mathcal{D}^m}\mathbb{1}_{[(S,T)\in B']}.\end{aligned}$$

Fix some S. Then, either $\mathbb{1}_{[S\in B]} = 0$ or $S \in B$ and then $\exists h_S$ such that $\mathcal{D}(h_S) \geq \epsilon$ and $|h_S \cap S| = 0$. It follows that a sufficient condition for $(S,T) \in B'$ is that $|T \cap h_S| > \frac{\epsilon m}{2}$. Therefore, whenever $S \in B$ we have

$$\mathop{\mathbb{E}}_{T\sim\mathcal{D}^m}\mathbb{1}_{[(S,T)\in B']} \;\geq\; \mathop{\mathbb{P}}_{T\sim\mathcal{D}^m}[|T\cap h_S| > \tfrac{\epsilon m}{2}].$$

But, since we now assume $S \in B$ we know that $\mathcal{D}(h_S) = \rho \geq \epsilon$. Therefore, $|T \cap h_S|$ is a binomial random variable with parameters ρ (probability of success for a single try) and m (number of tries). Chernoff's inequality implies

$$\mathbb{P}[|T\cap h_S| \leq \tfrac{\rho m}{2}] \;\leq\; e^{-\frac{2}{m\rho}(m\rho - m\rho/2)^2} = e^{-m\rho/2} \leq e^{-m\epsilon/2} \leq e^{-d\log(1/\delta)/2} = \delta^{d/2} \leq 1/2.$$

Thus,

$$\mathbb{P}[|T\cap h_S| > \tfrac{\epsilon m}{2}] \;=\; 1 - \mathbb{P}[|T\cap h_S| \leq \tfrac{\epsilon m}{2}] \;\geq\; 1 - \mathbb{P}[|T\cap h_S| \leq \tfrac{\rho m}{2}] \;\geq\; 1/2.$$

Combining all the preceding we *conclude the proof of Claim 1*.

Claim 2 (Symmetrization):
$\mathbb{P}[(S,T)\in B'] \leq e^{-\epsilon m/4}\,\tau_{\mathcal{H}}(2m)$.
Proof of Claim 2: To simplify notation, let $\alpha = m\epsilon/2$ and for a sequence $A = (x_1,\ldots,x_{2m})$ let $A_0 = (x_1,\ldots,x_m)$. Using the definition of B' we get that

$$\begin{aligned}\mathbb{P}[A\in B'] &= \mathop{\mathbb{E}}_{A\sim\mathcal{D}^{2m}}\max_{h\in\mathcal{H}}\mathbb{1}_{[\mathcal{D}(h)\geq\epsilon]}\mathbb{1}_{[|h\cap A_0|=0]}\mathbb{1}_{[|h\cap A|\geq\alpha]} \\ &\leq \mathop{\mathbb{E}}_{A\sim\mathcal{D}^{2m}}\max_{h\in\mathcal{H}}\mathbb{1}_{[|h\cap A_0|=0]}\mathbb{1}_{[|h\cap A|\geq\alpha]}.\end{aligned}$$

Now, let us define by $\mathcal{H}_A$ the effective number of different hypotheses on A, namely, $\mathcal{H}_A = \{h\cap A : h\in\mathcal{H}\}$. It follows that

$$\begin{aligned}\mathbb{P}[A\in B'] &\leq \mathop{\mathbb{E}}_{A\sim\mathcal{D}^{2m}}\max_{h\in\mathcal{H}_A}\mathbb{1}_{[|h\cap A_0|=0]}\mathbb{1}_{[|h\cap A|\geq\alpha]} \\ &\leq \mathop{\mathbb{E}}_{A\sim\mathcal{D}^{2m}}\sum_{h\in\mathcal{H}_A}\mathbb{1}_{[|h\cap A_0|=0]}\mathbb{1}_{[|h\cap A|\geq\alpha]}.\end{aligned}$$

Let $J = \{\mathbf{j}\subset[2m] : |\mathbf{j}| = m\}$. For any $\mathbf{j}\in J$ and $A = (x_1,\ldots,x_{2m})$ define $A_{\mathbf{j}} = (x_{j_1},\ldots,x_{j_m})$. Since the elements of A are chosen i.i.d., we have that for any $\mathbf{j}\in J$ and any function $f(A,A_0)$ it holds that $\mathbb{E}_{A\sim\mathcal{D}^{2m}}[f(A,A_0)] = \mathbb{E}_{A\sim\mathcal{D}^{2m}}[f(A,A_{\mathbf{j}})]$. Since this holds for any $\mathbf{j}$ it also holds for the expectation of $\mathbf{j}$ chosen at random from J. In particular, it holds for the function $f(A,A_0) = \sum_{h\in\mathcal{H}_A}\mathbb{1}_{[|h\cap A_0|=0]}\mathbb{1}_{[|h\cap A|\geq\alpha]}$. We

therefore obtain that

$$\begin{aligned}\mathbb{P}[A \in B'] &\leq \mathop{\mathbb{E}}_{A\sim\mathcal{D}^{2m}} \mathop{\mathbb{E}}_{j\sim J} \sum_{h\in\mathcal{H}_A} \mathbb{1}_{[|h\cap A_{\mathbf{j}}|=0]}\, \mathbb{1}_{[|h\cap A|\geq\alpha]} \\ &= \mathop{\mathbb{E}}_{A\sim\mathcal{D}^{2m}} \sum_{h\in\mathcal{H}_A} \mathbb{1}_{[|h\cap A|\geq\alpha]} \mathop{\mathbb{E}}_{j\sim J} \mathbb{1}_{[|h\cap A_{\mathbf{j}}|=0]}.\end{aligned}$$

Now, fix some A s.t. $|h \cap A| \geq \alpha$. Then, $\mathbb{E}_j \mathbb{1}_{[|h\cap A_{\mathbf{j}}|=0]}$ is the probability that when choosing m balls from a bag with at least α red balls, we will never choose a red ball. This probability is at most

$$(1-\alpha/(2m))^m = (1-\epsilon/4)^m \leq e^{-\epsilon m/4}.$$

We therefore get that

$$\mathbb{P}[A \in B'] \leq \mathop{\mathbb{E}}_{A\sim\mathcal{D}^{2m}} \sum_{h\in\mathcal{H}_A} e^{-\epsilon m/4} \leq e^{-\epsilon m/4} \mathop{\mathbb{E}}_{A\sim\mathcal{D}^{2m}} |\mathcal{H}_A|.$$

Using the definition of the growth function we *conclude the proof of Claim 2.*

Completing the Proof: By Sauer's lemma we know that $\tau_{\mathcal{H}}(2m) \leq (2em/d)^d$. Combining this with the two claims we obtain that

$$\mathbb{P}[S \in B] \leq 2(2em/d)^d e^{-\epsilon m/4}.$$

We would like the right-hand side of the inequality to be at most δ; that is,

$$2(2em/d)^d e^{-\epsilon m/4} \leq \delta.$$

Rearranging, we obtain the requirement

$$m \geq \frac{4}{\epsilon}\left(d\log(2em/d) + \log(2/\delta)\right) = \frac{4d}{\epsilon}\log(m) + \frac{4}{\epsilon}(d\log(2e/d) + \log(2/\delta).$$

Using Lemma A.2, a sufficient condition for the preceding to hold is that

$$m \geq \frac{16d}{\epsilon}\log\left(\frac{8d}{\epsilon}\right) + \frac{8}{\epsilon}(d\log(2e/d) + \log(2/\delta).$$

A sufficient condition for this is that

$$\begin{aligned}m &\geq \frac{16d}{\epsilon}\log\left(\frac{8d}{\epsilon}\right) + \frac{16}{\epsilon}(d\log(2e/d) + \tfrac{1}{2}\log(2/\delta) \\ &= \frac{16d}{\epsilon}\left(\log\left(\frac{8d\,2e}{d\epsilon}\right)\right) + \frac{8}{\epsilon}\log(2/\delta) \\ &= \frac{8}{\epsilon}\left(2d\log\left(\frac{16e}{\epsilon}\right) + \log\left(\frac{2}{\delta}\right)\right).\end{aligned}$$

and this concludes our proof. □

28.3.1 From ϵ-Nets to PAC Learnability

Theorem 28.4. *Let $\mathcal{H}$ be a hypothesis class over $\mathcal{X}$ with* $\mathrm{VCdim}(\mathcal{H}) = d$. *Let $\mathcal{D}$ be a distribution over $\mathcal{X}$ and let $c \in \mathcal{H}$ be a target hypothesis. Fix $\epsilon, \delta \in (0,1)$ and let m be as defined in Theorem 28.3. Then, with probability of at least $1-\delta$ over a choice of m*

i.i.d. instances from $\mathcal{X}$ with labels according to c we have that any ERM hypothesis has a true error of at most ϵ.

Proof. Define the class $\mathcal{H}^c = \{c \triangle h : h \in \mathcal{H}\}$, where $c \triangle h = (h \setminus c) \cup (c \setminus h)$. It is easy to verify that if some $A \subset \mathcal{X}$ is shattered by $\mathcal{H}$ then it is also shattered by $\mathcal{H}^c$ and vice versa. Hence, $\text{VCdim}(\mathcal{H}) = \text{VCdim}(\mathcal{H}^c)$. Therefore, using Theorem 28.3 we know that with probability of at least $1 - \delta$, the sample S is an ϵ-net for $\mathcal{H}^c$. Note that $L_{\mathcal{D}}(h) = \mathcal{D}(h \triangle c)$. Therefore, for any $h \in \mathcal{H}$ with $L_{\mathcal{D}}(h) \geq \epsilon$ we have that $|(h \triangle c) \cap S| > 0$, which implies that h cannot be an ERM hypothesis, which concludes our proof. □