Theorem 1 *Suppose algorithm A finds a hypothesis $h_A \in \mathcal{H}$ that is consistent with all N training examples (i.e., has training error zero). Then with probability at least $1-\delta$*

$$\mathrm{err}(h_A) \le \frac{\ln|\mathcal{H}| + \ln(1/\delta)}{N}.$$

Proof: Let

$$\epsilon = \frac{\ln|\mathcal{H}| + \ln(1/\delta)}{N},$$

and let us say that a hypothesis h is *ϵ-bad* if $\mathrm{err}(h) > \epsilon$. The goal is to show that h_A is *not* ϵ-bad (with probability at least $1-\delta$). That is, we want to show that

$$\Pr[h_A \text{ not } \epsilon\text{-bad}] \ge 1-\delta$$

or equivalently

$$\Pr[h_A \text{ is } \epsilon\text{-bad}] \le \delta.$$

We know that h_A is consistent with the training data. Thus,

$$\begin{aligned} \Pr[h_A \text{ is } \epsilon\text{-bad}] &= \Pr[h_A \text{ is consistent and } \epsilon\text{-bad}] \\ &\le \Pr[\exists h \in \mathcal{H} : h \text{ is consistent and } \epsilon\text{-bad}] \\ &= \Pr[\exists h \in \mathcal{B} : h \text{ is consistent}] \\ &= \Pr\left[h_1 \text{ consistent} \vee \cdots \vee h_{|\mathcal{B}|} \text{ consistent}\right] \\ &\le \Pr[h_1 \text{ consistent}] + \cdots + \Pr\left[h_{|\mathcal{B}|} \text{ consistent}\right]. \end{aligned}$$

Here, $\mathcal{B}$ is the set of all ϵ-bad hypotheses, which we list explicitly as $h_1, \ldots, h_{|\mathcal{B}|}$. That is,

$$\begin{aligned} \mathcal{B} &= \{h \in \mathcal{H} : h \text{ is } \epsilon\text{-bad}\} \\ &= \{h_1, \ldots, h_{|\mathcal{B}|}\}. \end{aligned}$$

Let h be any hypothesis in $\mathcal{B}$. Then

$$\begin{aligned} \Pr[h \text{ consistent}] &= \Pr[h(x_1) = f(x_1) \wedge \cdots \wedge h(x_N) = f(x_N)] \\ &= \Pr[h(x_1) = f(x_1)] \cdot \cdots \cdot \Pr[h(x_N) = f(x_N)] \\ &\le (1-\epsilon)^N. \end{aligned}$$

So, continuing the derivation above,

$$\begin{aligned} \Pr[h_A \text{ is } \epsilon\text{-bad}] &\le |\mathcal{B}| \cdot (1-\epsilon)^N \\ &\le |\mathcal{H}| \cdot (1-\epsilon)^N \\ &\le |\mathcal{H}| \cdot e^{-\epsilon N} \\ &= \delta. \end{aligned}$$

■