

Lecture 6: Agnostic Learning

Han Shao

Chapter 4 and Sections 6.4, 28.1, and 28.2, UML

So far, we have assumed realizability: there exists a hypothesis $h^* \in \mathcal{H}$ with

$$\text{err}_{\mathcal{D},f}(h^*) = 0.$$

Thus, the labels agree with h^* almost surely. In this lecture, we drop this assumption for two reasons:

1. There may be no deterministic labeling function, because of label noise or because the available features do not fully determine the label.
2. A deterministic ground-truth labeling function may exist but not belong to $\mathcal{H}$. In this case, our prior knowledge, as represented by the hypothesis class, is incorrect.

A joint distribution over instances and labels In the previous lectures, $\mathcal{D}$ was a distribution over $\mathcal{X}$, and labels were generated by a target function f . From now on, $\mathcal{D}$ denotes a *joint distribution* over $\mathcal{X} \times \mathcal{Y}$, and each labeled example (x, y) is drawn from this distribution.

Let $\mathcal{D}_{\mathcal{X}}$ denote the marginal distribution over instances. For any measurable set $B \subseteq \mathcal{X}$,

$$\mathcal{D}_{\mathcal{X}}(B) = \sum_{y \in \mathcal{Y}} \mathcal{D}(B \times \{y\}).$$

For an instance x , let $\mathcal{D}_{\mathcal{Y}|x}$ denote the conditional label distribution, so that

$$\mathcal{D}_{\mathcal{Y}|x}(y) = \Pr[Y = y \mid X = x],$$

where $(X, Y) \sim \mathcal{D}$. Sampling from the joint distribution can therefore be viewed as first drawing $x \sim \mathcal{D}_{\mathcal{X}}$ and then drawing $y \sim \mathcal{D}_{\mathcal{Y}|x}$. When $\mathcal{X}$ is discrete, this factorization reads

$$\Pr[X = x, Y = y] = \Pr[X = x] \Pr[Y = y \mid X = x].$$

This does not assume that instances and labels are independent: the conditional label distribution can depend on x .

True error, empirical error, and ERM For a hypothesis $h : \mathcal{X} \rightarrow \mathcal{Y}$, its *true error* is

$$\text{err}_{\mathcal{D}}(h) = \Pr_{(x,y) \sim \mathcal{D}}[h(x) \neq y].$$

There is no separate target function in this notation, since $\mathcal{D}$ already specifies how labels are generated.

Given a labeled training sample

$$S = ((x_1, y_1), \dots, (x_m, y_m)) \sim \mathcal{D}^m,$$

the *empirical error* is

$$\text{err}_S(h) = \frac{1}{m} \sum_{i=1}^m \mathbb{1}\{h(x_i) \neq y_i\}.$$

An empirical risk minimizer (ERM) returns

$$h_S \in \underset{h \in \mathcal{H}}{\text{argmin}} \text{err}_S(h).$$

Unlike in the realizable setting, the minimum empirical error need not be zero.

The Bayes optimal classifier The best possible prediction at a given instance is its most likely label. Thus, a *Bayes optimal classifier* is

$$h_{\text{Bayes}}(x) = \begin{cases} -1 & \text{if } \Pr[Y = -1 \mid X = x] \geq \frac{1}{2}, \\ +1 & \text{otherwise.} \end{cases}$$

Either label is optimal in a tie; the displayed convention chooses -1 . This classifier minimizes true error over all classifiers, but need not belong to the chosen class $\mathcal{H}$. Our learning guarantee will compare against the best error achievable *within* $\mathcal{H}$.

1 Agnostic PAC Learning

Definition 1 (Agnostic PAC Learnability). *A hypothesis class $\mathcal{H}$ is agnostic PAC learnable if there exist a function*

$$m_{\mathcal{H}} : (0, 1)^2 \rightarrow \mathbb{N}$$

and a learning algorithm $\mathcal{A}$ with the following property.

For every $\varepsilon, \delta \in (0, 1)$ and every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, when $\mathcal{A}$ receives $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ i.i.d. labeled examples from $\mathcal{D}$, it returns a hypothesis h_S such that, with probability at least $1 - \delta$,

$$\text{err}_{\mathcal{D}}(h_S) \leq \inf_{h \in \mathcal{H}} \text{err}_{\mathcal{D}}(h) + \varepsilon.$$

The difference

$$\text{err}_{\mathcal{D}}(h_S) - \inf_{h \in \mathcal{H}} \text{err}_{\mathcal{D}}(h)$$

is the *excess error* relative to $\mathcal{H}$. In the realizable setting, the infimum is zero, and the guarantee reduces to the one used in the previous lectures.

Theorem 1 (Finite classes are agnostically PAC learnable). *Let $\mathcal{H}$ be a finite, nonempty hypothesis class. For every $\varepsilon, \delta \in (0, 1)$, ERM returns a hypothesis satisfying*

$$\text{err}_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} \text{err}_{\mathcal{D}}(h) + \varepsilon$$

with probability at least $1 - \delta$, provided that

$$m \geq \frac{2}{\varepsilon^2} \left(\ln |\mathcal{H}| + \ln \frac{2}{\delta} \right).$$

The proof has the same basic structure as the finite-class proof in the realizable setting: control one fixed hypothesis, then apply a union bound. Here we use Hoeffding's inequality to control deviations between true and empirical error.

Hoeffding Bound Let $X_1, \dots, X_m$ be independent random variables taking values in $[0, 1]$ with common expectation μ . Then, for any $\eta > 0$,

$$\Pr \left[\left| \frac{1}{m} \sum_{i=1}^m X_i - \mu \right| \geq \eta \right] \leq 2 \exp(-2m\eta^2).$$

Proof of Theorem 1. It suffices to show that, with probability at least $1 - \delta$,

$$|\text{err}_S(h) - \text{err}_{\mathcal{D}}(h)| \leq \frac{\varepsilon}{2} \quad \text{for every } h \in \mathcal{H}.$$

This is a *uniform convergence* guarantee: the same sample accurately estimates the true error of every hypothesis simultaneously.

Why uniform convergence is enough. Let

$$h^* \in \underset{h \in \mathcal{H}}{\operatorname{argmin}} \text{err}_{\mathcal{D}}(h).$$

Such a minimizer exists because $\mathcal{H}$ is finite. Suppose the uniform convergence event holds, and consider any “bad” hypothesis h' with

$$\text{err}_{\mathcal{D}}(h') > \text{err}_{\mathcal{D}}(h^*) + \varepsilon.$$

Then

$$\text{err}_S(h') \geq \text{err}_{\mathcal{D}}(h') - \frac{\varepsilon}{2} > \text{err}_{\mathcal{D}}(h^*) + \frac{\varepsilon}{2} \geq \text{err}_S(h^*).$$

Thus h' cannot be an empirical risk minimizer. Every ERM output therefore has excess error at most ε .

Establishing uniform convergence. Fix $h \in \mathcal{H}$. The mistake indicators $X_i = \mathbb{1}\{h(x_i) \neq y_i\}$ are independent Bernoulli random variables with common expectation $\text{err}_{\mathcal{D}}(h)$. By Hoeffding’s inequality,

$$\Pr_{S \sim \mathcal{D}^m} \left[|\text{err}_S(h) - \text{err}_{\mathcal{D}}(h)| > \frac{\varepsilon}{2} \right] \leq 2 \exp \left(-\frac{m\varepsilon^2}{2} \right).$$

A union bound over $h \in \mathcal{H}$ yields

$$\begin{aligned} & \Pr_{S \sim \mathcal{D}^m} \left[\exists h \in \mathcal{H} : |\text{err}_S(h) - \text{err}_{\mathcal{D}}(h)| > \frac{\varepsilon}{2} \right] \\ & \leq 2|\mathcal{H}| \exp \left(-\frac{m\varepsilon^2}{2} \right) \leq \delta, \end{aligned}$$

where the final inequality follows from the sample-size condition. $\square$

For infinite classes, a direct union bound over all hypotheses is no longer useful. As in the realizable setting, we instead count the labelings induced on a finite sample and use Sauer’s Lemma.

Theorem 2 (VC-dimension upper bound; stated without proof). *Let $\mathcal{H}$ have finite VC dimension $d \geq 1$. There is a universal constant $C > 0$ such that, for all $\varepsilon, \delta \in (0, 1/2)$, ERM agnostically PAC learns $\mathcal{H}$ whenever*

$$m \geq C \frac{d + \ln(1/\delta)}{\varepsilon^2}.$$

A proof of this sharper upper bound uses a technique called *chaining*, which we do not cover here. Instead, by adopting the same double-sampling argument in the realizable setting, we could obtain a slightly looser bound

$$O \left(\frac{d \ln(1/\varepsilon) + \ln(1/\delta)}{\varepsilon^2} \right).$$

The Double-Sampling Argument Similar to the realizable setting, we could obtain the following growth-function bound.

Theorem 3 (Uniform convergence via the growth function). *Let $\varepsilon \in (0, 1)$. If*

$$m \geq \frac{8 \ln 4}{\varepsilon^2},$$

then

$$\Pr_{S \sim \mathcal{D}^m} \left[\exists h \in \mathcal{H} : |\text{err}_S(h) - \text{err}_{\mathcal{D}}(h)| > \frac{\varepsilon}{2} \right] \leq 4\Pi_{\mathcal{H}}(2m) \exp \left(-\frac{m\varepsilon^2}{32} \right). \quad (1)$$

Consequently, for every $\delta \in (0, 1)$, ERM has excess error at most ε with probability at least $1 - \delta$ whenever

$$m \geq \frac{32}{\varepsilon^2} \left(\ln(4\Pi_{\mathcal{H}}(2m)) + \ln \frac{1}{\delta} \right).$$

Proof. For $S \sim \mathcal{D}^m$, define the bad event

$$A_S = \left\{ \exists h \in \mathcal{H} : |\text{err}_S(h) - \text{err}_{\mathcal{D}}(h)| > \frac{\varepsilon}{2} \right\}.$$

By uniform convergence, our goal is to bound $\Pr_S[A_S]$.

Introducing a ghost sample. Draw an independent sample $S' \sim \mathcal{D}^m$ and define

$$B_{S,S'} = \left\{ \exists h \in \mathcal{H} : |\text{err}_S(h) - \text{err}_{S'}(h)| > \frac{\varepsilon}{4} \right\}.$$

This event compares two empirical errors instead of an empirical error and the unknown true error.

Claim 1. If $m \geq 8 \ln 4 / \varepsilon^2$, then

$$\Pr_{S,S'}[B_{S,S'}] \geq \frac{1}{2} \Pr_S[A_S].$$

Proof of Claim 1. Fix any sample S for which A_S occurs, and choose a witness $h \in \mathcal{H}$ satisfying

$$|\text{err}_S(h) - \text{err}_{\mathcal{D}}(h)| > \frac{\varepsilon}{2}.$$

Once S and the witness are fixed, S' is still an independent sample. By Hoeffding's inequality,

$$\Pr_{S'} \left[|\text{err}_{S'}(h) - \text{err}_{\mathcal{D}}(h)| > \frac{\varepsilon}{4} \right] \leq 2 \exp \left(-\frac{m\varepsilon^2}{8} \right) \leq \frac{1}{2}.$$

With probability at least $1/2$ over S' , the reverse triangle inequality therefore gives

$$\begin{aligned} |\text{err}_S(h) - \text{err}_{S'}(h)| &\geq |\text{err}_S(h) - \text{err}_{\mathcal{D}}(h)| - |\text{err}_{S'}(h) - \text{err}_{\mathcal{D}}(h)| \\ &> \frac{\varepsilon}{2} - \frac{\varepsilon}{4} = \frac{\varepsilon}{4}. \end{aligned}$$

Thus $\Pr_{S'}[B_{S,S'} \mid S] \geq 1/2$ for every such S . Averaging over S proves the claim. $\square$

An equivalent sampling process. Instead of drawing S and S' separately, we can:

1. Draw a pooled sample $S'' \sim \mathcal{D}^{2m}$.
2. Uniformly partition its $2m$ indexed sample occurrences into two samples S and S' , each of size m .

Repeated observations are treated as distinct sample occurrences. Define

$$B_{S'',S,S'}^* = \left\{ \exists h \in \mathcal{H} : |\text{err}_S(h) - \text{err}_{S'}(h)| > \frac{\varepsilon}{4} \right\}.$$

Claim 2. The two sampling processes have the same distribution for the events under consideration, so

$$\Pr_{S,S' \sim \mathcal{D}^m} [B_{S,S'}] = \Pr_{S'' \sim \mathcal{D}^{2m}, \text{partition}} [B_{S'',S,S'}^*].$$

Indeed, randomly permuting $2m$ i.i.d. examples and placing the first m in S and the remaining m in S' gives two independent i.i.d. samples.

Claim 3. For every fixed S'' and every fixed hypothesis h ,

$$\Pr_{\text{partition}} \left[|\text{err}_S(h) - \text{err}_{S'}(h)| > \frac{\varepsilon}{4} \mid S'' \right] \leq 2 \exp \left(-\frac{m\varepsilon^2}{32} \right).$$

Proof of Claim 3. Generate the random partition in the following equivalent way:

1. Randomly pair the $2m$ labeled sample occurrences:

$$(a_1, b_1), \dots, (a_m, b_m).$$

2. For each pair, independently draw a fair sign $\sigma_i \in \{-1, +1\}$. If $\sigma_i = +1$, send a_i to S and b_i to S' . If $\sigma_i = -1$, send a_i to S' and b_i to S .

For a labeled example $z = (x, y)$, write $\ell_h(z) = \mathbb{1}\{h(x) \neq y\}$. Condition on S'' and the pairing. Then

$$\text{err}_S(h) - \text{err}_{S'}(h) = \frac{1}{m} \sum_{i=1}^m \underbrace{\sigma_i (\ell_h(a_i) - \ell_h(b_i))}_{Z_i}.$$

The variables $Z_1, \dots, Z_m$ are independent, have mean zero, and lie in $[-1, 1]$. Their range has length 2.

Applying the general bounded-variable form of Hoeffding's inequality gives

$$\begin{aligned} \Pr_{\sigma} \left[\left| \frac{1}{m} \sum_{i=1}^m Z_i \right| > \frac{\varepsilon}{4} \mid S'', \text{pairing} \right] &\leq 2 \exp \left(-\frac{2m(\varepsilon/4)^2}{2^2} \right) \\ &= 2 \exp \left(-\frac{m\varepsilon^2}{32} \right). \end{aligned}$$

Equivalently, apply the $[0, 1]$ version to $(Z_i + 1)/2$, whose mean is $1/2$, at deviation $\varepsilon/8$. Since the bound holds for every pairing, averaging over the random pairing proves the claim. $\square$

A union bound over the induced labelings. Fix the pooled labeled sample

$$S'' = ((x_1'', y_1''), \dots, (x_{2m}'', y_{2m}'')),$$

and let $T'' = (x_1'', \dots, x_{2m}'')$ be its instance tuple. There are at most $\Pi_{\mathcal{H}}(2m)$ prediction vectors in $\mathcal{H}[T'']$. Because the observed labels in S'' are fixed, each prediction vector also determines all its mistake indicators on S'' . This remains true when the same instance appears with different observed labels.

Choose one representative hypothesis for each prediction vector and apply Claim 3. A union bound gives

$$\Pr_{\text{partition}} [B_{S'', S, S'}^* \mid S''] \leq 2\Pi_{\mathcal{H}}(2m) \exp\left(-\frac{m\varepsilon^2}{32}\right).$$

Averaging over S'' and combining Claims 1 and 2 now yields

$$\begin{aligned} \Pr_S [A_S] &\leq 2 \Pr_{S, S'} [B_{S, S'}] \\ &= 2 \Pr_{S'', \text{partition}} [B_{S'', S, S'}^*] \\ &\leq 4\Pi_{\mathcal{H}}(2m) \exp\left(-\frac{m\varepsilon^2}{32}\right). \end{aligned}$$

This proves (1). The theorem's final sample-size condition makes the last expression at most δ . It also implies $m \geq 8 \ln 4 / \varepsilon^2$, so Claim 1 applies. $\square$

Applying Sauer's Lemma For $d \geq 1$ and $2m > d + 1$, Sauer's Lemma gives

$$\Pi_{\mathcal{H}}(2m) \leq \left(\frac{2em}{d}\right)^d.$$

Consequently, the preceding argument bounds the failure probability by

$$4 \exp\left(d \ln \frac{2em}{d} - \frac{m\varepsilon^2}{32}\right).$$

For $\varepsilon, \delta \in (0, 1/2)$, taking

$$m \geq C \frac{d \ln(1/\varepsilon) + \ln(1/\delta)}{\varepsilon^2}$$

with a sufficiently large universal constant C makes this probability at most δ . Thus the double-sampling argument proves the claimed upper bound, with one additional $\ln(1/\varepsilon)$ factor in the VC-dimension term compared with theorem 2.

2 Sample-Complexity Lower Bound

Theorem 4 (Agnostic sample-complexity lower bound). *Let $\mathcal{H}$ be a binary hypothesis class with $1 \leq d = \text{VCDim}(\mathcal{H}) < \infty$. There are universal constants $c, \varepsilon_0, \delta_0 > 0$ such that, for every $\varepsilon \in (0, \varepsilon_0)$ and $\delta \in (0, \delta_0)$,*

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

We prove the two terms separately using a common coin-testing argument. For the VC-dimension term, we place a slightly biased coin at each point of a shattered set. For the confidence term, a single shattered point suffices.

Throughout this section, $\mathcal{V} = \{-1, +1\}$, and $h_S = \mathcal{A}(S)$ denotes the output of an arbitrary learning algorithm, which may be randomized and improper. The subscripts $S, \mathcal{A}$ include both the sample randomness and the learner's internal randomness; a subscript b additionally includes the random choice of hidden signs. The learner's random bits are independent of the data and the hidden signs. We use $\Pr_S$ for events determined solely by the sampled instances.

Distinguishing Two Biased Coins

Let $\rho \in (0, 1)$ be a bias parameter. An unknown coin is either

	Pr[head]	Pr[tail]
Coin A	$(1 + \rho)/2$	$(1 - \rho)/2$
Coin B	$(1 - \rho)/2$	$(1 + \rho)/2$

We observe N independent flips and try to determine which coin was selected.

An upper bound from Hoeffding's inequality. The majority rule chooses Coin A when more than half the flips are heads and Coin B when fewer than half are heads, breaking ties fairly. For $N \geq 1$, let K be the number of heads. Under Coin A, Hoeffding's inequality gives

$$\Pr_A \left[\frac{K}{N} \leq \frac{1}{2} \right] \leq \exp \left(-\frac{N\rho^2}{2} \right).$$

The corresponding upper-tail bound holds under Coin B. Hence $N = O(\ln(1/\delta)/\rho^2)$ flips suffice to identify the coin with error probability at most δ .

Lower bounds from Slud's inequality. Let $q_N(\rho)$ denote the minimum average error probability of any possibly randomized decision rule when the two coins are selected with equal probability. The following lemma supplies both versions of the testing lower bound that we will need.

Lemma 1 (Coin-testing lower bounds). *Let $0 < \rho \leq 1/2$ and let $N \geq 0$ be an integer. Then*

$$q_N(\rho) \geq \frac{1}{8} \exp(-2N\rho^2). \quad (2)$$

Moreover, if $N \leq 1/(2\rho^2)$, then

$$q_N(\rho) \geq c_0, \quad c_0 := \frac{1}{4} \left(1 - \sqrt{1 - e^{-1}} \right) > 0. \quad (3)$$

Proof. If $N = 0$, no information is available and $q_0(\rho) = 1/2$, so both bounds hold. Suppose $N \geq 1$. Comparing the likelihoods of a flip sequence under the two coins shows that the majority rule minimizes the average error. Ties have equal likelihood under the two coins. With fair tie-breaking, symmetry gives

$$q_N(\rho) = \Pr_B \left[K > \frac{N}{2} \right] + \frac{1}{2} \Pr_B \left[K = \frac{N}{2} \right] \geq \frac{1}{2} \Pr_B \left[K \geq \frac{N}{2} \right].$$

This optimality holds among randomized decision rules as well.

Under Coin B, $K \sim \text{Bin}(N, (1 - \rho)/2)$. Slud's inequality (UML, Appendix B.6, Lemma B.11), which we use without proof, gives

$$\Pr_B \left[K \geq \frac{N}{2} \right] \geq \frac{1}{2} \left(1 - \sqrt{1 - \exp \left(-\frac{N\rho^2}{1 - \rho^2} \right)} \right).$$

Consequently,

$$q_N(\rho) \geq \frac{1}{4} \left(1 - \sqrt{1 - \exp \left(-\frac{N\rho^2}{1 - \rho^2} \right)} \right). \quad (4)$$

For $u \in [0, 1]$,

$$1 - \sqrt{1 - u} = \frac{u}{1 + \sqrt{1 - u}} \geq \frac{u}{2}.$$

Applying this to (4), and using $1/(1 - \rho^2) \leq 2$, yields

$$q_N(\rho) \geq \frac{1}{8} \exp \left(-\frac{N\rho^2}{1 - \rho^2} \right) \geq \frac{1}{8} \exp(-2N\rho^2).$$

This proves (2).

If $N \leq 1/(2\rho^2)$, then

$$\frac{N\rho^2}{1 - \rho^2} \leq \frac{1}{2(1 - \rho^2)} \leq \frac{2}{3} < 1.$$

Substituting into (4) gives $q_N(\rho) \geq c_0$, proving (3). $\square$

A Family of Hard Distributions

Both lower bounds use the following construction. Let $T = \{x_1, \dots, x_k\}$ be a shattered set, and fix a bias $\rho \in (0, 1)$. For each $b = (b_1, \dots, b_k) \in \{-1, +1\}^k$, define $\mathcal{D}_b$ by

$$\Pr_{\mathcal{D}_b}[X = x_i] = \frac{1}{k}, \quad \Pr_{\mathcal{D}_b}[Y = y \mid X = x_i] = \begin{cases} (1 + \rho)/2 & \text{if } y = b_i, \\ (1 - \rho)/2 & \text{if } y = -b_i. \end{cases}$$

Thus b_i is the more likely label at x_i .

Since T is shattered, there is a hypothesis $h_b^* \in \mathcal{H}$ with $h_b^*(x_i) = b_i$ for every i . It predicts the more likely label at every point in the support, so

$$\inf_{h \in \mathcal{H}} \text{err}_{\mathcal{D}_b}(h) = \text{err}_{\mathcal{D}_b}(h_b^*) = \frac{1 - \rho}{2}.$$

Predicting $-b_i$ instead of b_i increases the conditional error by ρ . Therefore, the excess error of any classifier h is

$$W_b(h) := \text{err}_{\mathcal{D}_b}(h) - \text{err}_{\mathcal{D}_b}(h_b^*) = \frac{\rho}{k} \sum_{i=1}^k \mathbb{1}\{h(x_i) \neq b_i\}. \quad (5)$$

In particular, $0 \leq W_b(h) \leq \rho$, including when $h \notin \mathcal{H}$. We will use $k = d$ for the VC-dimension term and $k = 1$ for the confidence term.

Dependence on the VC Dimension

Proposition 1 (VC-dimension lower bound). *Let $\mathcal{H}$ have finite VC dimension $d \geq 1$, and let c_0 be the constant in lemma 1. For every $\varepsilon \in (0, c_0/8)$, every learning algorithm $\mathcal{A}$, and every integer sample size*

$$m \leq \frac{c_0^2 d}{64\varepsilon^2},$$

there exists a joint distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ such that

$$\Pr_{S, \mathcal{A}} \left[\text{err}_{\mathcal{D}}(h_S) > \inf_{h \in \mathcal{H}} \text{err}_{\mathcal{D}}(h) + \varepsilon \right] \geq \frac{c_0}{4}.$$

Proof. Fix $\mathcal{A}$ and set

$$\rho = \frac{4\varepsilon}{c_0}.$$

Then $0 < \rho < 1/2$, and the sample-size condition becomes $m \leq d/(4\rho^2)$. Use the construction above with a shattered set $T = \{x_1, \dots, x_d\}$ of size $k = d$, and choose b uniformly from $\{-1, +1\}^d$.

Few observations at each point. For $S = ((x'_1, y'_1), \dots, (x'_m, y'_m)) \sim \mathcal{D}_b^m$, let

$$N_i = \sum_{j=1}^m \mathbb{1}\{x'_j = x_i\}.$$

The instance marginal does not depend on b , and $\mathbb{E}_S[N_i] = m/d$. By Markov's inequality,

$$\Pr_S \left[N_i > \frac{1}{2\rho^2} \right] \leq 2\rho^2 \mathbb{E}_S[N_i] = \frac{2\rho^2 m}{d} \leq \frac{1}{2}.$$

Thus, for every $i \in [d]$,

$$\Pr_S \left[N_i \leq \frac{1}{2\rho^2} \right] \geq \frac{1}{2}. \tag{6}$$

A constant probability of predicting each sign incorrectly. Fix i and condition on all sampled instances and all labels observed at points other than x_i . The sign b_i remains uniform: the instance marginal is independent of b , and labels at other points depend only on the other independent signs.

If $N_i = n$, the labels observed at x_i are n independent flips of Coin A when $b_i = +1$ and Coin B when $b_i = -1$, with $+1$ interpreted as heads. The prediction $h_S(x_i)$ is a possibly randomized decision rule for this testing problem; the rest of the sample is fixed under the conditioning. By lemma 1, its average error is at least c_0 whenever $n \leq 1/(2\rho^2)$.

Averaging over the conditioned data gives, for every such n with $\Pr_S[N_i = n] > 0$,

$$\Pr_{b, S, \mathcal{A}} [h_S(x_i) \neq b_i \mid N_i = n] \geq c_0.$$

Together with (6), this implies

$$\Pr_{b, S, \mathcal{A}} [h_S(x_i) \neq b_i] \geq c_0 \Pr_S \left[N_i \leq \frac{1}{2\rho^2} \right] \geq \frac{c_0}{2}.$$

From sign errors to excess error. By (5) and linearity of expectation,

$$\begin{aligned}\mathbb{E}_{b,S,\mathcal{A}}[W_b(h_S)] &= \frac{\rho}{d} \sum_{i=1}^d \Pr_{b,S,\mathcal{A}}[h_S(x_i) \neq b_i] \\ &\geq \frac{c_0 \rho}{2} = 2\varepsilon.\end{aligned}$$

No independence between the learner's errors at different points is needed. Averaging over the finitely many vectors b gives a fixed b such that $\mathbb{E}_{S,\mathcal{A}}[W_b(h_S)] \geq 2\varepsilon$. This choice does not depend on the realized sample or the learner's random choices.

For this fixed b , write $W = W_b(h_S)$. Since $0 \leq W \leq \rho$,

$$\begin{aligned}2\varepsilon &\leq \mathbb{E}_{S,\mathcal{A}}[W] \\ &\leq \varepsilon \Pr_{S,\mathcal{A}}[W \leq \varepsilon] + \rho \Pr_{S,\mathcal{A}}[W > \varepsilon] \\ &\leq \varepsilon + \rho \Pr_{S,\mathcal{A}}[W > \varepsilon].\end{aligned}$$

Therefore,

$$\Pr_{S,\mathcal{A}}[W > \varepsilon] \geq \frac{\varepsilon}{\rho} = \frac{c_0}{4}.$$

By the definition of W_b , this proves the proposition with $\mathcal{D} = \mathcal{D}_b$. □

Dependence on the Confidence Parameter

Proposition 2 (Confidence lower bound). *Let $\mathcal{H}$ be a binary hypothesis class with $\text{VCDim}(\mathcal{H}) \geq 1$. For every $\varepsilon \in (0, 1/4]$, $\delta \in (0, 1/8)$, learning algorithm $\mathcal{A}$, and integer sample size*

$$m < \frac{\ln(1/(8\delta))}{8\varepsilon^2},$$

there exists a joint distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ such that

$$\Pr_{S,\mathcal{A}} \left[\text{err}_{\mathcal{D}}(h_S) > \inf_{h \in \mathcal{H}} \text{err}_{\mathcal{D}}(h) + \varepsilon \right] > \delta.$$

Proof. Fix $\mathcal{A}$. Since $\text{VCDim}(\mathcal{H}) \geq 1$, there is a shattered singleton $\{x_1\}$. Use the common construction with $k = 1$ and $\rho = 2\varepsilon \leq 1/2$, writing $b \in \{-1, +1\}$ for the hidden sign at x_1 . By (5),

$$W_b(h_S) = \rho \mathbb{1}\{h_S(x_1) \neq b\}.$$

Since $\rho > \varepsilon$, excess error greater than ε is equivalent to predicting b incorrectly.

Choose b uniformly. Every training example has instance x_1 , so $h_S(x_1)$ is a decision rule based on m flips of the corresponding coin. Applying (2),

$$\begin{aligned}\Pr_{b,S,\mathcal{A}}[W_b(h_S) > \varepsilon] &= \Pr_{b,S,\mathcal{A}}[h_S(x_1) \neq b] \\ &\geq q_m(\rho) \geq \frac{1}{8} \exp(-2m\rho^2) \\ &= \frac{1}{8} \exp(-8m\varepsilon^2) > \delta,\end{aligned}\tag{7}$$

where the final inequality uses the sample-size condition. At least one fixed sign b has failure probability at least this average. Taking $\mathcal{D} = \mathcal{D}_b$ for that sign proves the proposition. □

Combining the Two Lower Bounds

Proof of theorem 4. Set

$$\varepsilon_0 = \frac{c_0}{8}, \quad \delta_0 = \min \left\{ \frac{c_0}{4}, \frac{1}{64} \right\}.$$

For $\varepsilon \in (0, \varepsilon_0)$ and $\delta \in (0, \delta_0)$, both propositions apply. Moreover, $\ln(1/(8\delta)) \geq \frac{1}{2} \ln(1/\delta)$. Thus

$$m_{\mathcal{H}}(\varepsilon, \delta) \geq \max \left\{ \frac{c_0^2 d}{64\varepsilon^2}, \frac{\ln(1/\delta)}{16\varepsilon^2} \right\}.$$

Using $\max\{a, b\} \geq (a + b)/2$ and $0 < c_0 < 1$ gives

$$\begin{aligned} m_{\mathcal{H}}(\varepsilon, \delta) &\geq \frac{1}{2} \left(\frac{c_0^2 d}{64\varepsilon^2} + \frac{\ln(1/\delta)}{16\varepsilon^2} \right) \\ &\geq \frac{c_0^2}{128} \frac{d + \ln(1/\delta)}{\varepsilon^2}. \end{aligned}$$

Hence $c = c_0^2/128$ is a valid universal constant. □