Uniform Convergence via Symmetrization

0. Notation and Setup

- $S = (z_1, \ldots, z_m) \sim D^m$: i.i.d. sample
- $S' = (z'_1, \ldots, z'_m) \sim D^m$: ghost sample, independent of S
- $\ell(h, z) \in [0, 1]$: loss function
- $L_D(h) = \mathbb{E}_{z \sim D}[\ell(h, z)]$: true risk
- $L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)$: empirical risk
- $\tau_{\mathcal{H}}(2m)$: growth function

A. From True Risk to Difference of Two Samples

Step 1 (Markov motivation). Since $\sup_{h \in \mathcal{H}} |L_D(h) - L_S(h)| \geq 0$, Markov's inequality implies that bounding

$$\mathbb{E}_S \left[\sup_{h \in \mathcal{H}} |L_D(h) - L_S(h)| \right]$$

is sufficient for a high-probability bound.

Step 2 (Ghost sample identity). For every $h \in \mathcal{H}$,

$$L_D(h) = \mathbb{E}_{z \sim D}[\ell(h, z)] = \mathbb{E}_{S'} \left[\frac{1}{m} \sum_{i=1}^{m} \ell(h, z'_i) \right] = \mathbb{E}_{S'}[L_{S'}(h)].$$

Additional explanation

0. The Setting

Data are generated i.i.d. from an unknown true distribution D. We select a hypothesis h from a hypothesis class $\mathcal{H}$, and our goal is to guarantee that the empirical risk $L_S(h)$ computed from a training sample is close to the true risk $L_D(h)$, *simultaneously for all hypotheses* in $\mathcal{H}$.

1. What Is a Single Data Point z?

Definition.

$$\boxed{z = (x, y)}$$

- x: input (feature, problem instance)
- y: ground-truth label

A data point is *not* just the label. It is a single row consisting of both the input and the corresponding label.

Why not only the label? The loss function typically has the form

$$\ell(h, z) = \ell(h, (x, y)),$$

which requires both the prediction $h(x)$ and the true label y. Therefore, z must include both components.

Concrete Examples

Example A: Spam classification.

- x: "free coupon click" (email content)
- y: 1 (spam)

$$\boxed{z = (\text{“free coupon click”}, 1)}$$

Example B: Study time → pass/fail.

- x: study time (hours)
- y: pass = 1, fail = 0

If a student studied for 3 hours and failed,

$$\boxed{z = (x = 3, y = 0)}.$$

2. What Is a Hypothesis h?

Definition.

$$\boxed{h : \mathcal{X} \to \{0, 1\}}$$

A hypothesis is a rule (model) that maps an input x to a prediction $\hat{y} = h(x)$.

Examples

- Threshold classifier:

$$h_a(x) = \mathbf{1}[x < a]$$

- Linear classifier:

$$h_w(x) = \mathbf{1}[w^\top x \geq 0]$$

- Study-time rule:

$$\boxed{h(x) = \mathbf{1}[x \geq 5]}$$

"Predict pass if study time is at least 5 hours."

3. What Is the Loss $\ell(h, z)$?

Meaning.

$$\ell(h, z) \equiv \ell(h, (x, y))$$

It measures how wrong hypothesis h is on data point z.

Most basic choice: 0–1 loss.

$$\boxed{\ell(h, (x, y)) = \mathbf{1}[h(x) \neq y]}$$

- correct prediction $\Rightarrow 0$
- incorrect prediction $\Rightarrow 1$
- $\ell(h, z) \in [0, 1]$, which enables Hoeffding-type bounds

Explicit Computations

Study-time example. Let

$$h(x) = \mathbf{1}[x \geq 5], \qquad z = (x = 3, y = 1).$$

- prediction: $h(3) = 0$
- true label: $y = 1$

Thus,

$$\ell(h, z) = \mathbf{1}[0 \neq 1] = 1.$$

If instead $z = (x = 3, y = 0)$, then the prediction is correct and

$$\ell(h, z) = 0.$$

4. The Sample S and Empirical Risk $L_S(h)$

Training sample.

$$\boxed{S = (z_1, \ldots, z_m)}, \qquad z_i = (x_i, y_i).$$

Empirical risk.

$$\boxed{L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)}$$

This is the average loss of h on the training data.

Numerical Example

Let

$$S = \{(2,0),(6,1),(4,0)\}, \qquad h(x) = \mathbf{1}[x \geq 5].$$

- $z_1 = (2,0)$: $h(2) = 0$, correct $\Rightarrow \ell = 0$
- $z_2 = (6,1)$: $h(6) = 1$, correct $\Rightarrow \ell = 0$
- $z_3 = (4,0)$: $h(4) = 0$, correct $\Rightarrow \ell = 0$

Hence,

$$L_S(h) = \frac{1}{3}(0+0+0) = 0.$$

The hypothesis appears perfect on the sample.

5. The Distribution D and True Risk $L_D(h)$

Assume data are generated from an unknown true distribution D:

$$z \sim D.$$

True risk.

$$\boxed{L_D(h) = \mathbb{E}_{z\sim D}[\ell(h,z)]}$$

This represents the expected error on future data drawn from D. It is the *true generalization performance*.

6. Why Do $L_S(h)$ and $L_D(h)$ Differ?

- $L_S(h)$ is an average over a *finite* sample
- $L_D(h)$ is an expectation over an *infinite* population

With a small sample size, sampling noise causes fluctuations: easy samples make $L_S(h)$ small, difficult ones make it large.

Coin-Flipping Analogy

Even if the true probability of heads is 0.5:

- 7 heads out of 10 flips $\Rightarrow$ sample mean 0.7
- 3 heads out of 10 flips $\Rightarrow$ sample mean 0.3

A finite sample always fluctuates.

7. "Not Equal" vs. "Equal in Expectation"

Incorrect statement.

$$L_S(h) = L_D(h)$$

Correct statement.

$$\boxed{\mathbb{E}_S[L_S(h)] = L_D(h)}$$

That is, the *expectation* of the empirical risk equals the true risk.

Why This Is True

From

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i),$$

taking expectation yields

$$\mathbb{E}_S[L_S(h)] = \mathbb{E}_S \left[\frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i) \right] = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_S[\ell(h, z_i)].$$

Since each z_i is drawn i.i.d. from D,

$$\mathbb{E}_S[\ell(h, z_i)] = \mathbb{E}_{z \sim D}[\ell(h, z)] = L_D(h).$$

Thus,

$$\mathbb{E}_S[L_S(h)] = \frac{1}{m} \sum_{i=1}^{m} L_D(h) = L_D(h).$$

8. A Concrete Over-Optimism Example

Suppose

$$L_D(h) = 0.3,$$

meaning the hypothesis truly makes mistakes 30% of the time.

If we sample only $m = 10$ points and happen to observe just one mistake,

$$L_S(h) = 0.1.$$

The hypothesis appears much better than it actually is. As m increases, typically $L_S(h) \to L_D(h)$, but we want this convergence to hold *uniformly for all* h.

9. The Goal: Uniform Convergence

Our main quantity of interest is

$$\boxed{\sup_{h\in\mathcal{H}} \big|L_D(h) - L_S(h)\big|}$$

This measures the maximum deviation between empirical and true risk over the entire hypothesis class.

Formally, we aim to show

$$\Pr\left(\sup_{h\in\mathcal{H}} |L_D(h) - L_S(h)| \le \varepsilon\right) \ge 1 - \delta.$$

This property is called *uniform convergence*.

10. Why Symmetrization Starts with $L_D(h) = \mathbb{E}_{S'}[L_{S'}(h)]$

The difficulty is that $L_D(h)$ depends on the unknown distribution D.

Introduce an independent ghost sample

$$S' = (z'_1, \dots, z'_m) \sim D^m.$$

Then,

$$\boxed{L_D(h) = \mathbb{E}_{z\sim D}[\ell(h,z)] = \mathbb{E}_{S'}\left[\frac{1}{m}\sum_{i=1}^{m} \ell(h, z'_i)\right] = \mathbb{E}_{S'}[L_{S'}(h)]}$$

This replaces an expectation over the unknown distribution with an expectation over an independent sample average.

As a result:

- Original problem: $L_D(h)$ vs. $L_S(h)$ (distribution vs. sample)
- After symmetrization: $L_{S'}(h)$ vs. $L_S(h)$ (sample vs. sample)

This is the core symmetrization trick.

11. Summary

- $z = (x, y)$: a single data point (input + label)
- h: prediction rule
- $\ell(h, z)$: loss comparing prediction and truth
- $L_S(h)$: empirical (sample) risk
- $L_D(h)$: true (population) risk

- $L_S(h) \neq L_D(h)$ in general (sampling noise)
- $\mathbb{E}[L_S(h)] = L_D(h)$
- Goal: $\sup_{h\in\mathcal{H}} |L_D(h) - L_S(h)|$ small (uniform convergence)
- Symmetrization key step: $L_D(h) = \mathbb{E}_{S'}[L_{S'}(h)]$

Step 3 (Substitution). Substituting into the expectation,

$$\mathbb{E}_S\left[\sup_h |L_D(h) - L_S(h)|\right] = \mathbb{E}_S\left[\sup_h |\mathbb{E}_{S'}[L_{S'}(h)] - L_S(h)|\right].$$

Step 4 (Absolute value vs expectation). Using $|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$,

$$|\mathbb{E}_{S'}[L_{S'}(h) - L_S(h)]| \leq \mathbb{E}_{S'}[|L_{S'}(h) - L_S(h)|].$$

Jensen's inequality. The function $f(x) = |x|$ is convex. By Jensen's inequality,

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

Substituting $f(x) = |x|$ yields

$$|\mathbb{E}[X]| \leq \mathbb{E}[|X|].$$

Step 5 (Supremum vs expectation). Since $\sup_h \mathbb{E}[X_h] \leq \mathbb{E}[\sup_h X_h]$,

$$\sup_h \mathbb{E}_{S'}[|L_{S'}(h) - L_S(h)|] \leq \mathbb{E}_{S'}\left[\sup_h |L_{S'}(h) - L_S(h)|\right].$$

Why it holds. For every fixed hypothesis $h \in \mathcal{H}$ and for every realization of S',

$$|L_{S'}(h) - L_S(h)| \;\leq\; \sup_{g\in\mathcal{H}} |L_{S'}(g) - L_S(g)|.$$

Taking expectation $\mathbb{E}_{S'}[\cdot]$ on both sides preserves the inequality:

$$\mathbb{E}_{S'}[|L_{S'}(h) - L_S(h)|] \;\leq\; \mathbb{E}_{S'}\left[\sup_{g\in\mathcal{H}} |L_{S'}(g) - L_S(g)|\right].$$

Step 6 (Symmetrization inequality). Combining Steps 3–5,

$$\mathbb{E}_S\left[\sup_h |L_D(h) - L_S(h)|\right] \leq \mathbb{E}_{S,S'}\left[\sup_h |L_{S'}(h) - L_S(h)|\right]. \tag{1}$$

Note that

$$L_{S'}(h) - L_S(h) = \frac{1}{m}\sum_{i=1}^{m}\big(\ell(h, z_i') - \ell(h, z_i)\big).$$

B. Symmetrization and Concentration

Step 7 (Distributional symmetry). Since (z_i, z_i') are i.i.d., exchanging z_i and z_i' does not change the joint distribution.

1. Basic Objects

- A single data point:
$$z = (x, y)$$
- A hypothesis (classifier):
$$h : \mathcal{X} \to \{0, 1\}$$
- Loss function (fixed to 0–1 loss in this proof):
$$\ell(h, z) = \mathbf{1}[h(x) \neq y] \in \{0, 1\}$$
- Samples:
$$S = (z_1, \ldots, z_m), \qquad S' = (z_1', \ldots, z_m'),$$
both drawn i.i.d. from the distribution D

2. The Quantity We Truly Want to Control

$$\sup_{h \in \mathcal{H}} \big| L_D(h) - L_S(h) \big|$$

- $L_S(h)$: average loss on a finite sample
- $L_D(h)$: average loss under the true distribution

Core difficulty. The distribution D is unknown; we only observe samples.

3. First Trick: Symmetrization

Key identity.
$$L_D(h) = \mathbb{E}_{S'}[L_{S'}(h)]$$

Hence,
$$\mathbb{E}_S \left[\sup_h |L_D(h) - L_S(h)| \right] \;\leq\; \mathbb{E}_{S,S'} \left[\sup_h |L_{S'}(h) - L_S(h)| \right].$$

4. Writing the Sample Difference as a Sum

$$L_{S'}(h) - L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \big(\ell(h, z_i') - \ell(h, z_i) \big).$$

Thus, the core object becomes

$$\mathbb{E}_{S,S'}\left[\sup_h \frac{1}{m}\Big|\sum_{i=1}^m (\ell(h,z_i') - \ell(h,z_i))\Big|\right].$$

Step 8 (Sign flipping).

Key expression.

$$\mathbb{E}_{S,S'}\left[\sup_h \frac{1}{m}\Big|\sum_{i=1}^m (\ell(h,z_i') - \ell(h,z_i))\Big|\right]$$

has the *same distribution* as

$$\mathbb{E}_{S,S'}\left[\sup_h \frac{1}{m}\Big|\sum_{i=1}^m \sigma_i(\ell(h,z_i') - \ell(h,z_i))\Big|\right],$$

for any $\sigma \in \{\pm 1\}^m$.

Why this is valid.

- (z_i, z_i') are i.i.d.
- Therefore (z_i, z_i') and (z_i', z_i) have the same distribution
- Swapping the order flips the sign of

$$\ell(h,z_i') - \ell(h,z_i)$$

Hence,

$$X_i := \ell(h,z_i') - \ell(h,z_i) \quad \text{and} \quad -X_i$$

have the same distribution. Multiplying by arbitrary signs $\sigma_i \in \{\pm 1\}$ preserves the joint distribution.

If an expression is identical for all sign vectors σ, it is also identical when averaged over a uniformly random σ:

$$\mathbb{E}_{\sigma\sim U_\pm^m}\mathbb{E}_{S,S'}\left[\sup_h \frac{1}{m}\Big|\sum_{i=1}^m \sigma_i(\ell(h,z_i') - \ell(h,z_i))\Big|\right].$$

This is a purely algebraic step, not a probabilistic assumption.

Step 9 (Rademacher averaging). For every fixed $\sigma \in \{\pm 1\}^m$, the distributional symmetry implies

$$\mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^m \big(\ell(h,z_i') - \ell(h,z_i)\big)\right|\right] = \mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i\big(\ell(h,z_i') - \ell(h,z_i)\big)\right|\right].$$

Averaging over a uniformly random σ yields

$$\mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m}\big(\ell(h,z_i')-\ell(h,z_i)\big)\right|\right]=\mathbb{E}_{\sigma}\mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i\big(\ell(h,z_i')-\ell(h,z_i)\big)\right|\right].$$

Since (S,S') and σ are independent, by Fubini's theorem,

$$\mathbb{E}_{\sigma}\mathbb{E}_{S,S'}\left[\cdots\right]=\mathbb{E}_{S,S'}\mathbb{E}_{\sigma}\left[\cdots\right].$$

That is,

$$\mathbb{E}_{\sigma}\mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i\big(\ell(h,z_i')-\ell(h,z_i)\big)\right|\right]=\mathbb{E}_{S,S'}\mathbb{E}_{\sigma}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i\big(\ell(h,z_i')-\ell(h,z_i)\big)\right|\right].$$

At this point, we view (S,S') as fixed and analyze the inner expectation only with respect to the randomness of σ.

Meaning of "Fixing (S,S')"

Precise meaning. At this stage, (S,S') are treated as fixed constants, and only the randomness of σ is considered:

$$\mathbb{E}_{S,S'}\big[\mathbb{E}_{\sigma}[\cdot\mid S,S']\big].$$

Why this matters. To apply Hoeffding's inequality, we need:

- independent random variables
- bounded support by deterministic constants

The bounds

$$\sigma_i(\ell(h,z_i')-\ell(h,z_i))\in[-1,1]$$

hold only once (z_i,z_i') are fixed.

Step 10 (Finite reduction). Fix (S,S') and let C be the set of $2m$ points appearing in them. Then

$$\sup_{h\in\mathcal{H}}(\cdot)=\max_{h\in\mathcal{H}_C}(\cdot),\quad |\mathcal{H}_C|\le\tau_{\mathcal{H}}(2m).$$

Set of observed inputs.

$$C:=\{x_i\}\cup\{x_i'\},\qquad |C|\le 2m.$$

Key observation. The hypothesis h appears only through its values on C:

$$h(x_i), \quad h(x_i').$$

Thus, hypotheses are equivalent if they induce the same labeling on C.

Definition.

$$\mathcal{H}_C := \{\text{distinct labelings of } C \text{ induced by } \mathcal{H}\}.$$

Hence,

$$\sup_{h \in \mathcal{H}}(\cdot) = \max_{h \in \mathcal{H}_C}(\cdot),$$

reducing an infinite set to a finite one.

Step 11 (Definition of θ_h). For $h \in \mathcal{H}_C$, define

$$\theta_h = \frac{1}{m} \sum_{i=1}^{m} \sigma_i (\ell(h, z_i') - \ell(h, z_i)).$$

Why Hoeffding applies. For fixed (h, S, S'):

- σ_i are independent
- $\ell(h, z) \in \{0, 1\}$
- $\ell(h, z_i') - \ell(h, z_i) \in \{-1, 0, 1\}$
- thus $\sigma_i(\ell(h, z_i') - \ell(h, z_i)) \in [-1, 1]$
- $\mathbb{E}_\sigma[\theta_h] = 0$ (symmetry implies mean zero.)

If $\text{VCdim}(\mathcal{H}) = d$, Sauer's lemma gives

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

After symmetrization and Rademacher randomization, we arrive at the following setting.

- $S, S' \sim D^m$ are independent i.i.d. samples.
- $\sigma_1, \ldots, \sigma_m$ are i.i.d. Rademacher variables, uniform on $\{\pm 1\}$.
- C is the set of inputs appearing in (S, S'), with $|C| \leq 2m$.
- $\mathcal{H}_C$ is the finite set of distinct labelings of C induced by $\mathcal{H}$.

For fixed (S, S'), define

$$\theta_h := \frac{1}{m}\sum_{i=1}^{m} \sigma_i\big(\ell(h, z_i') - \ell(h, z_i)\big), \qquad h \in \mathcal{H}_C.$$

For the 0–1 loss, $\ell \in \{0, 1\}$ and hence $\ell(h, z_i') - \ell(h, z_i) \in \{-1, 0, 1\} \subset [-1, 1]$.

Step 12 (Hoeffding). Conditioned on (S, S'), the random variables $\sigma_i(\ell(h, z_i') - \ell(h, z_i))$ are independent, mean zero, and bounded in $[-1, 1]$. Thus, Hoeffding's inequality gives, for any $\rho > 0$,

$$\Pr_{\sigma}\big(|\theta_h| > \rho\big) \leq 2e^{-2m\rho^2}. \tag{1}$$

Step 13 (Union bound). Since we are interested in $\max_{h\in\mathcal{H}_C} |\theta_h|$, applying a union bound yields

$$\Pr_{\sigma}\Big(\max_{h\in\mathcal{H}_C} |\theta_h| > \rho\Big) \leq \sum_{h\in\mathcal{H}_C} \Pr_{\sigma}(|\theta_h| > \rho) \leq 2|\mathcal{H}_C|e^{-2m\rho^2}. \tag{2}$$

Define the nonnegative random variable

$$X := \max_{h\in\mathcal{H}_C} |\theta_h|.$$

Then it implies that for all $\rho > 0$,

$$\Pr(X > \rho) \leq 2|\mathcal{H}_C|e^{-2m\rho^2}. \tag{3}$$

Step 14 (Tail integration). Integrating the tail bound (Lemma A.4),

$$\mathbb{E}\Big[\max_{h\in\mathcal{H}_C} |\theta_h|\Big] \leq \frac{4 + \sqrt{\log|\mathcal{H}_C|}}{\sqrt{2m}}.$$

Lemma A.4 (Tail-to-Expectation Conversion)

Let $X \geq 0$ be a random variable such that

$$\Pr(X > t) \leq 2b\, e^{-t^2/a^2} \quad \text{for all } t > 0.$$

Then

$$\mathbb{E}[X] \leq a\big(2 + \sqrt{\log b}\big).$$

Applying Lemma A.4

Comparing with the lemma, we identify

$$b := |\mathcal{H}_C|, \qquad a := \frac{1}{\sqrt{2m}}.$$

Therefore,

$$\mathbb{E}_\sigma\Big[\max_{h\in\mathcal{H}_C}|\theta_h|\Big] \le \frac{1}{\sqrt{2m}}\Big(2+\sqrt{\log|\mathcal{H}_C|}\Big). \tag{4}$$

In many texts, constants are loosened for simplicity, leading to the equivalent bound

$$\mathbb{E}_\sigma\Big[\max_{h\in\mathcal{H}_C}|\theta_h|\Big] \le \frac{4+\sqrt{\log|\mathcal{H}_C|}}{\sqrt{2m}}. \tag{4'}$$

The key scaling is

$$\mathbb{E}[\max_h|\theta_h|] = O\left(\sqrt{\frac{\log|\mathcal{H}_C|}{m}}\right).$$

Since C contains at most $2m$ points,

$$|\mathcal{H}_C| \le \tau_{\mathcal{H}}(|C|) \le \tau_{\mathcal{H}}(2m). \tag{5}$$

Step 15 (Growth function bound).

$$|\mathcal{H}_C| \le \tau_{\mathcal{H}}(2m).$$

Therefore,

$$\boxed{\mathbb{E}_S\Big[\sup_{h\in\mathcal{H}}|L_D(h)-L_S(h)|\Big] \le \frac{4+\sqrt{\log\tau_{\mathcal{H}}(2m)}}{\sqrt{2m}}}.$$

Substituting into (4'),

$$\mathbb{E}_{S,S',\sigma}\Big[\max_{h\in\mathcal{H}_C}|\theta_h|\Big] \le \frac{4+\sqrt{\log\tau_{\mathcal{H}}(2m)}}{\sqrt{2m}}. \tag{6}$$

Tracing back through symmetrization yields

$$\boxed{\mathbb{E}_S\Big[\sup_{h\in\mathcal{H}}|L_D(h)-L_S(h)|\Big] \le \frac{4+\sqrt{\log\tau_{\mathcal{H}}(2m)}}{\sqrt{2m}}}. \tag{7}$$

This is the expectation-form uniform convergence bound.

Let

$$X := \sup_{h\in\mathcal{H}}|L_D(h)-L_S(h)| \ge 0.$$

By Markov's inequality,

$$\Pr(X \ge t) \le \frac{\mathbb{E}[X]}{t}. \tag{8}$$

Setting $t = \mathbb{E}[X]/\delta$ gives

$$\Pr\Big(X \le \frac{\mathbb{E}[X]}{\delta}\Big) \ge 1-\delta. \tag{9}$$

Combining with (7),

$$\boxed{\sup_{h\in\mathcal{H}} |L_D(h) - L_S(h)| \le \frac{4 + \sqrt{\log \tau_{\mathcal{H}}(2m)}}{\delta\sqrt{2m}}} \quad \text{with probability at least } 1-\delta. \tag{10}$$

If $\mathrm{VCdim}(\mathcal{H}) = d$, then for $2m \ge d$,

$$\tau_{\mathcal{H}}(2m) \le \left(\frac{2em}{d}\right)^d. \tag{11}$$

Thus,

$$\log \tau_{\mathcal{H}}(2m) \le d \log\left(\frac{2em}{d}\right). \tag{12}$$

Substituting into (10),

$$\boxed{\sup_{h\in\mathcal{H}} |L_D(h) - L_S(h)| \le \frac{4 + \sqrt{d\log(2em/d)}}{\delta\sqrt{2m}}} \quad \text{w.p. } \ge 1-\delta. \tag{13}$$

We want

$$\sup_h |L_D(h) - L_S(h)| \le \varepsilon \quad \text{with probability } \ge 1-\delta.$$

A sufficient condition from (13) is

$$\frac{4 + \sqrt{d\log(2em/d)}}{\delta\sqrt{2m}} \le \varepsilon. \tag{14}$$

This leads to inequalities of the form

$$m \ge a \log m + b, \tag{15}$$

where

$$a = \frac{2d}{(\delta\varepsilon)^2}, \qquad b = \frac{2d\log(2e/d)}{(\delta\varepsilon)^2}.$$

Lemma A.2 (Solving $x \ge a\log x + b$)

If

$$x \ge 4a\log(2a) + 2b,$$

then

$$x \ge a\log x + b.$$

Applying this lemma yields the closed-form sufficient condition

$$\boxed{m \ge \frac{8d}{(\delta\varepsilon)^2}\log\Big(\frac{4d}{(\delta\varepsilon)^2}\Big) + \frac{4d\log(2e/d)}{(\delta\varepsilon)^2}}. \tag{16}$$

Why Assuming $\sqrt{d\log(2em/d)} \geq 4$ Is WLOG

Let

$$x := \sqrt{d\log(2em/d)}.$$

If $x \geq 4$, then

$$4 + x \leq 2x,$$

and the bound (13) simplifies to

$$\frac{1}{\delta}\sqrt{\frac{2d\log(2em/d)}{m}}. \tag{17}$$

If $x < 4$, then

$$4 + x < 8,$$

and a simpler bound

$$\frac{8}{\delta\sqrt{2m}} \tag{18}$$

applies.

Conclusion (Sample Complexity for VC Classes). Let $\mathcal{H}$ be a hypothesis class with $\mathrm{VCdim}(\mathcal{H}) = d$. To learn with accuracy ε and confidence $1 - \delta$, it suffices to take the sample size

$$\boxed{m = O\left(\frac{d}{\varepsilon^2}\log\frac{d}{\varepsilon\delta}\right).}$$