Topic 5: Uniform convergence

5.1 Uniform convergence of training error rates

5.1.1 Finite hypothesis classes

Our approach to analyzing ERM in Theorem 4.4 was to establish *uniform convergence* of training error rates to population error rates for hypotheses from $\mathcal{H}$: for $\mathbf{S} \sim P^T$ for any probability distribution P over $\mathcal{X} \times \mathcal{Y}$,

$$\Pr\left(\forall h \in \mathcal{H},\ |\mathsf{err}(h;\mathbf{S}) - \mathsf{err}(h;P)| \leq \sqrt{\frac{\ln(2|\mathcal{H}|/\delta)}{T}}\right) \geq 1-\delta. \tag{5.1}$$

In the $1-\delta$ probability event from (5.1), a bound is established on the difference between training and population error rates that holds uniformly for all $h \in \mathcal{H}$. This should be contrasted with a statement of the following form:

$$\forall h \in \mathcal{H},\ \Pr\left(|\mathsf{err}(h;\mathbf{S}) - \mathsf{err}(h;P)| \leq \sqrt{\frac{\ln(2/\delta)}{T}}\right) \geq 1-\delta. \tag{5.2}$$

In each of the $1-\delta$ probability events from (5.2), the difference between training and population error rates is bounded only for a single hypothesis. Going from (5.2) to (5.1) was just a matter of applying the union bound, replacing δ in (5.2) with $\delta/|\mathcal{H}|$. However, this is only possible if $\mathcal{H}$ is finite; for infinite $\mathcal{H}$, the simple union bound does not work.

5.1.2 Impossibility for some infinite hypothesis classes

For some infinite hypothesis classes, there is no hope to establish a uniform bound on $|\mathsf{err}(h;\mathbf{S}) - \mathsf{err}(h;P)|$ that tends to zero with the number of training data. For example, suppose $\mathcal{X} = [0,1]$, $\mathcal{Y} = \{0,1\}$, and $\mathcal{H} = \{h_V \mid V \subseteq \mathcal{X}, |V| < \infty\}$, where $h_V(x) := \mathbb{1}\{x \in V\}$ for $x \in \mathcal{X}$. And suppose $(\mathbf{x},\mathbf{y}) \sim P$ has $\mathbf{x}$ uniformly distributed on $[0,1]$, and $\mathbf{y} = 0$ deterministically. Then for all $h_V \in \mathcal{H}$,

$$\mathsf{err}(h_V;P) = \Pr(h_V(\mathbf{x}) = 1) = \int_0^1 \mathbb{1}\{x \in V\}\,\mathrm{d}x = 0.$$

However, with $\mathbf{S} = ((\mathbf{x}_1,\mathbf{y}_1),\ldots,(\mathbf{x}_T,\mathbf{y}_T)) \sim P^T$ and $\mathbf{V} := \{\mathbf{x}_1,\ldots,\mathbf{x}_T\}$, we have

$$\mathsf{err}(h_{\mathbf{V}};\mathbf{S}) = 1.$$

Therefore, we always have

$$\exists h_V \in \mathcal{H} \text{ s.t. } |\mathsf{err}(h_V;\mathbf{S}) - \mathsf{err}(h_V;P)| = 1.$$

What makes the above hypothesis class unsuitable for uniform convergence? Are there infinite hypothesis classes for which a uniform convergence guarantee like (5.1) holds?

5.2 Vapnik-Chervonenkis theory

5.2.1 Empirical covers and empirical processes

Vapnik and Chervonenkis (1971) developed a theory to study algorithms for learning binary classifiers, and a key part of their theory is a technique for proving uniform convergence of training error rates for hypothesis classes. Their basic technique has two components. The first component characterizes which functions classes can be effectively replaced by a finite approximation. The relevant finite approximation is called an *empirical cover*, and there is a combinatorial property of a function class that determines if it always admits a non-trivial empirical cover.

The second component of the technique concerns how to effectively use a "union bound" over an empirical cover to establish uniform convergence of empirical means over an iid sample to the population means. The (random) probability distribution $\mathbf{P}_n$ that puts $1/n$ mass on each point in a size-n iid sample from P is called an *empirical distribution* (or *empirical measure*). (The average value of a function under $\mathbf{P}_n$ is called an *empirical average*.) The study of how $\mathbf{P}_n$ deviates from P is the subject of empirical process theory, and Vapnik and Chervonenkis drawn up techniques from that literature.

5.2.2 VC dimension

Let $\mathcal{F} \subseteq \{0,1\}^{\mathcal{Z}}$ be a class of $\{0,1\}$-valued functions defined on a domain $\mathcal{Z}$. The *behaviors* of $\mathcal{F}$ on a finite set of m points $\{z_1, \ldots, z_m\} \subseteq \mathcal{Z}$ is

$$\mathcal{F}[\{z_1, \ldots, z_m\}] := \{(f(z_1), \ldots, f(z_m)) \mid f \in \mathcal{F}\} \subseteq \{0,1\}^m.$$

Intuitively, we only care about the behaviors of $\mathcal{F}$ on a finite set of points when it comes to understanding the empirical averages of functions from $\mathcal{F}$. However, the number of behaviors, even on a finite set, can vary a lot depending on the function class. What about a function class determines whether the number of behaviors is small or large?

We say that $\mathcal{F}$ *shatters* the m points $\{z_1, \ldots, z_m\}$ if

$$\mathcal{F}[\{z_1, \ldots, z_m\}] = \{0,1\}^m;$$

that is, for each $(b_1, \ldots, b_m) \in \{0,1\}^m$, there exists some $f \in \mathcal{F}$ such that $f(z_i) = b_i$ for all $i \in [m]$. The *Vapnik-Chervonenkis (VC) dimension* of $\mathcal{F}$, denoted $\mathsf{VCdim}(\mathcal{F})$, is the largest value m such that there exists a set of m points from $\mathcal{Z}$ that is shattered by $\mathcal{F}$; if there are arbitrarily largest sets of point shattered by $\mathcal{F}$, then $\mathsf{VCdim}(\mathcal{F}) = \infty$.

Lemma 5.1. *For any $\mathcal{F} \subseteq \{0,1\}^{\mathcal{Z}}$, $\mathsf{VCdim}(\mathcal{F}) \leq \log_2(|\mathcal{F}|)$.*

Proof. If a set $\{z_1, \ldots, z_m\}$ of cardinality m is shattered by $\mathcal{F}$, then $|\mathcal{F}[\{z_1, \ldots, z_m\}]| = 2^m$. Now map every $f \in \mathcal{F}$ to $(f(z_1), \ldots, f(z_m)) \in \mathcal{F}[\{z_1, \ldots, z_m\}]$ to deduce $|\mathcal{F}| \geq 2^m$ as well. □

Example 5.1. Recall the class of monotone disjunctions MDISJ_n over the Boolean n-cube $\{0,1\}^n$:

$$\mathsf{MDISJ}_n := \{d_S \mid S \subseteq [n]\}$$

where $d_S(x_1, \ldots, x_n) = \bigvee_{j \in S} x_j$. We claim that $\mathsf{VCdim}(\mathsf{MDISJ}_n) = n$. We have $\mathsf{VCdim}(\mathsf{MDISJ}_n) \leq n$ by Lemma 5.1 since $|\mathsf{MDISJ}_n| = 2^n$. So it remains to show that $\mathsf{VCdim}(\mathsf{MDISJ}_n) \geq n$.

We claim that the set $T := \{e_1, \ldots, e_n\}$ is shattered by MDISJ_n; here, $e_i \in \{0,1\}^n$ is the vector with a 1 in the i-th position and 0s elsewhere. To show this, we must establish that every possible labeling of T can be realized by functions from MDISJ_n. Consider an arbitrary labeling $b \in \{0,1\}^n$, written $b = (b_1, \ldots, b_n)$; b_i is the desired label of e_i. Set $S := \{j \in [n] \mid b_j = 1\}$. Then for each $i \in [n]$,

$$d_S(e_i) = \bigvee_{j \in S} e_i[j] = \bigvee_{j \in S} \mathbb{1}\{i = j\} = \mathbb{1}\{i \in S\} = b_i.$$

Therefore the labeling b is realized by the monotone disjunction d_S. We conclude that T is shattered by MDISJ_n, and hence $\mathsf{VCdim}(\mathsf{MDISJ}_n) \geq |T| = n$. $\Diamond$

Example 5.2. Recall the class of linear threshold functions over Euclidean n-space $\mathbb{R}^n$:

$$\mathsf{LTF}_{\mathbb{R}^n} := \Big\{ f \colon \mathbb{R}^n \to \{0,1\} \;\Big|\; \exists (w, \theta) \in \mathbb{R}^n \times \mathbb{R} \text{ s.t. } f(x) = \mathbb{1}\{\langle w, x \rangle > \theta\} \Big\}.$$

We claim that $\mathsf{VCdim}(\mathsf{LTF}_{\mathbb{R}^n}) = n + 1$.

We first show $\mathsf{VCdim}(\mathsf{LTF}_{\mathbb{R}^n}) \geq n + 1$ by exhibiting a set of $n + 1$ points that is shattered by $\mathsf{LTF}_{\mathbb{R}^n}$. We claim that the set $T := \{0, e_1, \ldots, e_n\}$, where e_i is the i-th coordinate basis vector in $\mathbb{R}^n$, is shattered by $\mathsf{LTF}_{\mathbb{R}^n}$. To show this, we must establish that every possible labeling of T can be realized by functions from $\mathsf{LTF}_{\mathbb{R}^n}$. Consider an arbitrary labeling $b \in \{0,1\}^{n+1}$, written $b = (b_0, b_1, \ldots, b_n)$; b_0 is the desired label of 0, and for each $i \in [n]$, b_i is the desired label of e_i. Let $\tilde{b} := 2b - 1 = (\tilde{b}_0, \tilde{b}_1, \ldots, \tilde{b}_n) \in \{-1, 1\}^{n+1}$ be the $\{-1,1\}$-version of b, and consider the weight vector w and threshold θ given by

$$w := 2\tilde{b}_1 e_1 + \cdots + 2\tilde{b}_n e_n, \quad \theta := -\tilde{b}_0$$

Then

$$\mathbb{1}\{\langle w, 0 \rangle > \theta\} = \mathbb{1}\{0 > -\tilde{b}_0\} = b_0,$$

and for each $i \in [n]$,

$$\mathbb{1}\{\langle w, e_i \rangle > \theta\} = \mathbb{1}\{2\tilde{b}_i > -\tilde{b}_0\} = \mathbb{1}\{\tilde{b}_i + \tilde{b}_0/2 > 0\} = b_i.$$

Therefore the labeling b is realized by the LTF $x \mapsto \mathbb{1}\{\langle w, x \rangle > \theta\}$. We conclude that T is shattered by $\mathsf{LTF}_{\mathbb{R}^n}$, and hence $\mathsf{VCdim}(\mathsf{LTF}_{\mathbb{R}^n}) \geq |T| = n + 1$.

It now remains to show that $\mathsf{VCdim}(\mathsf{LTF}_{\mathbb{R}^n}) < n + 2$. To do this, we use the following simple observation (recalling the discussion in Section 3.3.3 about Perceptron): if a set $T \subset \mathbb{R}^n$ is shattered by $\mathsf{LTF}_{\mathbb{R}^n}$, then $\tilde{T} := \{(x, -1) \mid x \in T\} \subset \mathbb{R}^{n+1}$ is shattered by $\mathsf{HLTF}_{\mathbb{R}^{n+1}}$. So it suffices to show that no set of $n + 2$ points is shattered by $\mathsf{HLTF}_{\mathbb{R}^{n+1}}$. Consider any $n + 2$ points $x_1, \ldots, x_{n+2} \in \mathbb{R}^{n+1}$. They must be linearly dependent, so we can write

$$c_1 x_1 + \cdots + c_{n+2} x_{n+2} = 0$$

for some $c_1, \ldots, c_{n+1} \in \mathbb{R}$ that are not all zero. In particular,

$$\langle w, c_1 x_1 + \cdots + c_{n+2} x_{n+2} \rangle = 0, \quad \forall w \in \mathbb{R}^{n+1}. \tag{5.3}$$

We may assume without loss of generality that $c_i > 0$ for some $i \in [n+2]$ (or else replace all c_i with $-c_i$). Define $b = (b_1, \ldots, b_{n+2}) \in \{0,1\}^{n+2}$ by $b_i := \mathbb{1}\{c_i > 0\}$ for each $i \in [n+2]$. Suppose, for sake

of contradiction, that the labeling b is realized by some HLTF, i.e., there exists $w \in \mathbb{R}^{n+1}$ such that, for all $i \in [n+2]$,

$$\langle w, x_i \rangle > 0 \iff c_i > 0.$$

This implies

$$\left\langle w, \sum_{i=1}^{n+2} c_i x_i \right\rangle = \sum_{i=1}^{n+2} c_i \langle w, x_i \rangle > 0,$$

which is a contradiction of (5.3). So the labeling b of $x_1, \ldots, x_{n+2}$ cannot be realized by any HLTF. Since this holds for any choice of $n+2$ points from $\mathbb{R}^{n+1}$, it follows that no set of $n+2$ points from $\mathbb{R}^{n+1}$ is shattered by $\mathsf{HLTF}_{\mathbb{R}^{n+1}}$. This means that no set of $n+2$ points from $\mathbb{R}^n$ is shattered by $\mathsf{LTF}_{\mathbb{R}^n}$, and hence $\mathsf{VCdim}(\mathsf{LTF}_{\mathbb{R}^n}) < n+2$. ◊

Exercise 5.1. Let $\mathcal{X} := \{0, 1, 2, \ldots, N-1\}$ for $N \in \mathbb{N}$, and

$$\mathcal{H} := \{h \colon \mathcal{X} \to \{0,1\} \mid \exists \theta \in \mathbb{R} \text{ s.t. } h(x) = \mathbb{1}\{x > \theta\} \; \forall x \in \mathcal{X}\}.$$

What are $\mathsf{Ldim}(\mathcal{H})$ and $\mathsf{VCdim}(\mathcal{H})$? ◊

Exercise 5.2. Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ denote a hypothesis class on the domain $\mathcal{X}$. Let $\mathcal{H}^* := \{\mathsf{e}_x \mid x \in \mathcal{X}\}$ be the dual class where

$$\mathsf{e}_x(h) := h(x) \quad \forall h \in \mathcal{H}.$$

(a) Show that

$$\mathsf{VCdim}(\mathcal{H}^*) \leq 2^{\mathsf{VCdim}(\mathcal{H})+1} - 1.$$

(b) Let $\mathcal{X} := \mathbb{R}^n$ and $\mathcal{H} := \{h \colon \mathbb{R}^n \to \{0,1\} \mid \exists w \in \mathbb{R}^n \text{ s.t. } h(x) = \mathbb{1}\{\langle w, x \rangle > 0\} \; \forall x \in \mathbb{R}^n\}$. Show that $\mathsf{VCdim}(\mathcal{H}^*) = \mathsf{VCdim}(\mathcal{H})$. ◊

5.2.3 Sauer's lemma

If $\mathcal{F}$ has VC dimension d, then (by definition) there are finite point sets $T \subset \mathcal{Z}$ of all sizes $|T| \leq d$ for which the behaviors of $\mathcal{F}$ on T is all behaviors: $|\mathcal{F}[T]| = 2^n$. However, if $d < \infty$, this exponential growth stops as soon as $|T| \geq d$, after which the number of behaviors grows at most polynomially.

Define the *growth function* $\Pi_{\mathcal{F}} \colon \mathbb{N} \to \mathbb{N}$ for the function class $\mathcal{F}$ by

$$\Pi_{\mathcal{F}}(n) := \max\big\{|\mathcal{F}[T]| \;\big|\; T \subseteq \mathcal{Z}, |T| = n\big\}, \quad \forall n \in \mathbb{N}.$$

Sauer's lemma, given in Lemma 5.2, implies that $\Pi_{\mathcal{F}}$ is asymptotically upper-bounded by a polynomial in n whenever $\mathsf{VCdim}(\mathcal{F}) < \infty$.

Lemma 5.2 (Sauer, 1972; Shelah, 1972)**.** *If* $d := \mathsf{VCdim}(\mathcal{F}) < \infty$*, then* $\Pi_{\mathcal{F}}(n) \leq \binom{n}{\leq d} := \sum_{k=0}^{d} \binom{n}{k}$.

There are a number of proofs of Lemma 5.2. The proof we give below is based on those of Alon (1983) and Frankl (1983) using a technique called shifting.

Lemma 5.2 concerns subsets of the Boolean n-cube $\{0,1\}^n$, so it is useful to specialize the notions of shattering and VC dimension to such subsets. For a subset $V \subseteq \{0,1\}^n$ and $J \subseteq [n]$, let $V|_J := \{(v_i)_{i \in J} \mid v \in V\}$. We say V shatters J if $V|_J = \{0,1\}^J$, and define $\mathsf{VCdim}(V)$ to be the largest value d such that there exists $J \subseteq [n]$ of cardinality $|J| = d$ such that V shatters T.

Note that each $v \in \{0,1\}^n$ can also be regarded as the characteristic vector for a subset of $[n]$; we will write v to also refer to that subset. With this overloaded notation, we have that $V \subseteq 2^{[n]}$ shatters $J \subseteq [n]$ if and only if $\{v \cap J \mid v \in V\} = 2^J$. When v is regarded this way, we also use the shorthand $v - i$ to mean $v \setminus \{i\}$ for any $i \in [n]$.

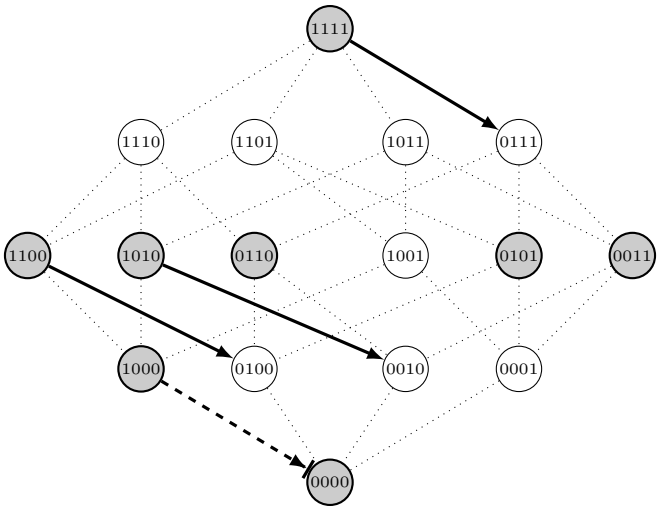


Figure 5.1: Shifting $V \subseteq \{0,1\}^n$ (for $n = 4$) along coordinate $i = 1$. The family V is indicated by the gray colored vertices. (Vertices corresponding to $v \in \{0,1\}^n \setminus V$ are also depicted but colored white.) The dark solid arrows indicate $v \in V$ that are successfully shifted (e.g., $\mathsf{S}_{1,V}((1,1,0,0)) = (0,1,0,0)$). Successful shifts move $v \in V$ down to lower levels of the Boolean n-cube. The dark dashed arrow indicates an unsuccessful attempt to shift $v = (1,0,0,0)$ that is blocked by the presence of $(0,0,0,0) \in V$.

Proposition 5.1. *For any finite set $T \subseteq \mathcal{Z}$, we have $\mathsf{VCdim}(\mathcal{F}[T]) \leq \mathsf{VCdim}(\mathcal{F})$.*

Exercise 5.3. Prove Proposition 5.1, and explain in what sense the inequality is tight. ◊

The goal of the proof of Lemma 5.2 is to perform some transformations to $V := \mathcal{F}[T]$ (for some $T \subseteq \mathcal{Z}$ with $|T| = n$) that (i) preserve the cardinality of V, and (ii) do not increase the VC dimension of V. After all of the transformations, the resulting family $W \subseteq 2^{[n]}$ will have some useful properties that make it very easy to bound $|W|$ (and hence also $|V|$).

The transformations to V are called *(down)shifting* along a coordinate $i \in [n]$. The shifting transformations are as follows.

$$\text{Shift } v \in V \text{ along coordinate } i \in [n]: \qquad \mathsf{S}_{i,V}(v) := \begin{cases} v - i & \text{if } i \in v \text{ and } v - i \notin V, \\ v & \text{otherwise.} \end{cases}$$

$$\text{Shift } V \text{ along coordinate } i \in [n]: \qquad \mathsf{S}_i(V) := \{\mathsf{S}_{i,V}(v) \mid v \in V\}.$$

If $\mathsf{S}_{i,V}(v) = v - i \neq v$, then we say that v is successfully shifted along coordinate i. See Figure 5.1 for an illustration. There are two possible reasons for an unsuccessful shift along coordinate i. First, it could be that $i \notin v$. Second, it could be that $i \in v$ but $v - i \in V$; in this case, the attempted shift is "blocked" by the presence of $v - i \in V$.

Exercise 5.4. For the $V \subseteq \{0,1\}^4$ from Figure 5.1, list the members of $\mathsf{S}_2(V)$. ◊

Proposition 5.2. *For any $V \subseteq 2^{[n]}$ and any $i \in [n]$:*

1. $|\mathsf{S}_i(V)| = |V|$.

2. *For any $J \subseteq [n]$, if J is shattered by $\mathsf{S}_i(V)$, then J is shattered by V. (Consequently, $\mathsf{VCdim}(\mathsf{S}_i(V)) \leq \mathsf{VCdim}(V)$.)*

Proof. Observe that $v \in V$ is successfully shifted by $\mathsf{S}_{i,V}$ if and only if $v - i \neq V$; so V and $\mathsf{S}_i(V)$ are in 1-to-1 correspondence (via $\mathsf{S}_{i,V}$). This proves the first item of the claim.

To prove the second item, consider any $J \subseteq [n]$ shattered by $\mathsf{S}_i(V)$, so $\{\mathsf{S}_{i,V}(v) \cap J \mid v \in V\} = 2^J$. We need to show that J is also shattered by V. There are two cases to consider.

- Case 1: $i \notin J$. Then for every $v \in V$, we have $\mathsf{S}_{i,V}(v) \cap J = v \cap J$ since $\mathsf{S}_{i,V}$ only (possibly) removes i from v. So $\{v \cap J \mid v \in V\} = \{\mathsf{S}_{i,V}(v) \cap J \mid v \in V\} = 2^J$, which shows that J is also shattered by V.

- Case 2: $i \in J$. Since $\{\mathsf{S}_{i,V}(v) \cap J \mid v \in V\} = 2^J$, for any choice of $K \in 2^J$, there exists $v \in V$ such that
$$\mathsf{S}_{i,V}(v) \cap J = K \cup \{i\}.$$
This means that $i \in \mathsf{S}_{i,V}(v)$, implying that $\mathsf{S}_{i,V}(v) = v$, an unsuccessful shift of $v \in V$ along coordinate i, blocked by the presence of $u := v - i$ in V. Therefore, we have both
$$u \cap J = K \setminus \{i\} \in V \quad \text{and}$$
$$v \cap J = K \cup \{i\} \in V.$$
Since this holds for all choices of $K \subseteq J$, it follows that V also shatters J. □

Exercise 5.5. Can there exist $V \subseteq \{0,1\}^n$ such that $\mathsf{VCdim}(\mathsf{S}_1(V)) < \mathsf{VCdim}(V)$? ◊

Suppose, starting with V, we repeatedly apply S_i to V for each $i \in [n]$. Whenever there is a successful shift, the "total weight" of V (i.e., $\sum_{v \in V} |v|$) decreases by at least 1. So this process must eventually stop with some set $W \subseteq 2^{[n]}$ with $|W| = |V|$ and $\mathsf{VCdim}(W) \leq \mathsf{VCdim}(V)$.

Proposition 5.3. *Suppose $W \subseteq 2^{[n]}$ satisfies $\mathsf{S}_i(W) = W$ for each $i \in [n]$. For each $v \in W$, we have $u \in W$ for all $u \subseteq v$ (i.e., W is downward-closed).*

Proof. Pick any nonempty $v \in W$ and any $i \in v$. If $v - i \notin W$, then $\mathsf{S}_{i,W}(v)$ would be a successful shift, a contradiction of the assumption that $\mathsf{S}_i(W) = W$. Therefore it must be that $v - i \in W$. Repeatedly applying this argument shows that every subset $u \subseteq v$ is also in W. □

Proposition 5.4. *Suppose $W \subseteq 2^{[n]}$ is downward-closed, and let $d := \mathsf{VCdim}(W)$. Then*
$$|W| \leq \binom{n}{\leq d}.$$

Proof. We first claim that every $v \in W$ has $|v| \leq d$. To see this, consider any $v \in W$, and let $J := v$. Since W is downward-closed, W shatters J. Since $\mathsf{VCdim}(W) = d$, it must be that $|J| \leq d$. This establishes $|v| \leq d$.

So, every $v \in W$ is a subset of $[n]$ of cardinality at most d; and there are at most $\binom{n}{\leq d}$ subsets of $[n]$ of cardinality at most d. □

Combining Propositions 5.1 to 5.4 (along with the above observation that the number of successful shifts is finite) completes the proof Lemma 5.2.

5.2.4 Population and empirical averages

Let P be a probability distribution over $\mathcal{Z}$ (which we regard as the "population"), and let $\mathbf{z}_{1:n} := (\mathbf{z}_1, \dots, \mathbf{z}_n) \sim P^n$ be a size-n iid sample from P. For a real-valued function $f\colon \mathcal{Z} \to \mathbb{R}$, we are concerned with two different "averages":

- the *population average of f*
$$\mathrm{E}[f(\mathbf{z})]$$
(where $\mathbf{z} \sim P$);
- the *empirical average of f*
$$\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}_i).$$

Our goal is to show the following theorem.

Theorem 5.1. *Let $\mathcal{F} \subseteq \{0,1\}^{\mathcal{Z}}$ be a function class with $d := \mathsf{VCdim}(\mathcal{F}) < \infty$. Let P be a probability distribution over $\mathcal{Z}$ and $(\mathbf{z}_1, \dots, \mathbf{z}_n) \sim P^n$. Then*

$$\mathrm{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}_i) - \mathrm{E}[f(\mathbf{z})]\right|\right] \le 2\sqrt{\frac{2\ln(2\binom{n}{\le d})}{n}}.$$

A proof that has a lot of overlap with that of Theorem 5.1 establishes the following theorem.

Theorem 5.2 (Vapnik and Chervonenkis, 1971)**.** *Let $\mathcal{F} \subseteq \{0,1\}^{\mathcal{Z}}$ be a function class with $d := \mathsf{VCdim}(\mathcal{F}) < \infty$. Let P be a probability distribution over $\mathcal{Z}$ and $(\mathbf{z}_1, \dots, \mathbf{z}_n) \sim P^n$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}_i) - \mathrm{E}[f(\mathbf{z})]\right| \le O\left(\sqrt{\frac{\log(\binom{n}{\le d}/\delta)}{n}}\right).$$

It is also possible to derive Theorem 5.2 as an immediate corollary of Theorem 5.1 by applying a probability inequality called McDiarmid's inequality. We focus on Theorem 5.1 because its proof is more streamlined than the direct proof of Theorem 5.2 but retains most of the key ideas.

In Theorems 5.1 and 5.2, because $\mathcal{F}$ has finite VC dimension, we might hope that we can replace the supremum over $\mathcal{F}$ with a maximum over the set of "behaviors" of $\mathcal{F}$ on $\{\mathbf{z}_1, \dots, \mathbf{z}_n\}$. The thought would be that (at least for Theorem 5.2) because there are only $\binom{n}{\le d}$ behaviors, we can establish a probability inequality for each behavior separately, and then apply a union bound. This is almost how the proofs of Theorems 5.1 and 5.2 work, but some important issues arise that require careful attention.

The first issue is that, even though $\mathcal{F}$ has only finitely-many behaviors on the size-n iid sample—and hence finitely-many values of the empirical average $\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}_i)$ as f ranges over $\mathcal{F}$—there may be infinitely-many values of the population average $\mathrm{E}[f(\mathbf{z})]$ as f ranges over $\mathcal{F}$.

A way to resolve this issue is based on the following observation: if $\mathbf{x}$ and $\mathbf{x}'$ are iid real-valued random variables, then $\mathrm{E}[\mathbf{x}] = \mathrm{E}[\mathbf{x}' \mid \mathbf{x}]$, which implies $\mathrm{E}[|\mathbf{x} - \mathrm{E}[\mathbf{x}]|] \le \mathrm{E}[|\mathbf{x} - \mathbf{x}'|]$.[1] This suggests

[1] In fact, if we considered squared differences instead of absolute differences, the left-hand side would be equal to half the right-hand side: $\mathrm{E}[(\mathbf{x} - \mathrm{E}[\mathbf{x}])^2] = \frac{1}{2}\mathrm{E}[(\mathbf{x} - \mathbf{x}')^2]$.

that, for the purposes of analysis only, it may be useful to consider a second size-n iid sample $\mathbf{z}'_{1:n} := (\mathbf{z}'_1, \dots, \mathbf{z}'_n) \sim P^n$ that is independent of—but identically distributed as—the original sample $\mathbf{z}_{1:n}$. We call this second sample the *ghost sample* 👻. If we could replace $\mathrm{E}[f(\mathbf{z})]$ with the empirical average of f on the ghost sample, then we could again hope to use the fact that the (ghost) empirical average takes only finitely-many values as f ranges over $\mathcal{F}$. (It is still not clear how to exactly take advantage of this, but intuitively this could seem like progress.)

Proposition 5.5 (Symmetrization by ghost sample)**.** *Let $\mathbf{z}_{1:n}$ and $\mathbf{z}'_{1:n}$ be independent size-n iid samples from a probability distribution P over $\mathcal{Z}$. For any $\mathcal{F} \subseteq \{0,1\}^{\mathcal{Z}}$,*

$$\mathrm{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}_i) - \mathrm{E}[f(\mathbf{z})]\right|\right] \le \mathrm{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}_i) - \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}'_i)\right|\right].$$

Proof. Define

$$\boldsymbol{\Delta} f := \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}_i) - \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}'_i)$$

to be the difference between the two empirical averages of $f\colon \mathcal{Z} \to \mathbb{R}$. Consider any $f \in \mathcal{F}$, and observe that

$$\mathrm{E}[f(\mathbf{z})] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}_i)\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}'_i) \,\middle|\, \mathbf{z}_{1:n}\right]$$

since $\mathbf{z}_{1:n}$ and $\mathbf{z}'_{1:n}$ are independent and identically distributed. Therefore,

$$\left|\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}_i) - \mathrm{E}[f(\mathbf{z})]\right| = \left|\mathrm{E}\left[\boldsymbol{\Delta} f \,\middle|\, \mathbf{z}_{1:n}\right]\right| \le \mathrm{E}\left[|\boldsymbol{\Delta} f| \,\middle|\, \mathbf{z}_{1:n}\right]$$

where the inequality follows by Jensen's inequality ($|\mathrm{E}[\mathbf{x}]| \le \mathrm{E}[|\mathbf{x}|]$). We also have

$$|\boldsymbol{\Delta} f| \le \sup_{f'\in\mathcal{F}} |\boldsymbol{\Delta} f'|,$$

and therefore

$$\mathrm{E}\left[|\boldsymbol{\Delta} f| \,\middle|\, \mathbf{z}_{1:n}\right] \le \mathrm{E}\left[\sup_{f'\in\mathcal{F}} |\boldsymbol{\Delta} f'| \,\middle|\, \mathbf{z}_{1:n}\right].$$

Since $f \in \mathcal{F}$ was arbitrary, we have

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}_i) - \mathrm{E}[f(\mathbf{z})]\right| \le \mathrm{E}\left[\sup_{f'\in\mathcal{F}} |\boldsymbol{\Delta} f'| \,\middle|\, \mathbf{z}_{1:n}\right].$$

Taking expectation of both sides gives the claim. □

Exercise 5.6. In the same setting as Proposition 5.5, show that

$$\mathrm{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}_i) - \mathrm{E}[f(\mathbf{z})]\right|\right] \ge \frac{1}{2}\,\mathrm{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}_i) - \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}'_i)\right|\right]. \qquad \lozenge$$

Proposition 5.5 is called a "symmetrization" argument, because it effectively replaces each (possibly) asymmetric random variable $f(\mathbf{z}_i) - \mathrm{E}[f(\mathbf{z})]$ with a symmetric random variable $f(\mathbf{z}_i) - f(\mathbf{z}'_i)$. So, to prove Theorem 5.1, it suffices bound the right-hand side in Proposition 5.5. (In view of Exercise 5.6, at most a constant factor is lost.) But this raises the second issue: how to take advantage of the finitely-many behaviors of $\mathcal{F}$ on the (two) iid samples. Note that the set of behaviors is itself a random subset of $\{0,1\}^{2n}$; what leverage is gotten by fixing this set of behaviors?

The expression on the right-hand side in Proposition 5.5 concerns two independent iid samples, $\mathbf{z}_{1:n}$ and $\mathbf{z}'_{1:n}$, which we can imagine are derived by starting with a single size-$2n$ iid sample, say, $(\mathbf{z}^a_1, \mathbf{z}^b_1, \ldots, \mathbf{z}^a_n, \mathbf{z}^b_n) \sim P^{2n}$, and then, for each $i \in [n]$, assigning one of $\mathbf{z}^a_i$ and $\mathbf{z}^b_i$ to the original sample, and then assigning the other to the ghost sample. Any way of making this assignment is valid as long as it does not depend on the size-$2n$ iid sample. This flexibility turns out to be key.

Proposition 5.6. *Let $\mathbf{z}_{1:n}$ and $\mathbf{z}'_{1:n}$ be independent size-n iid samples from a probability distribution P over $\mathcal{Z}$. For any $\mathcal{F} \subseteq \{0,1\}^{\mathcal{Z}}$ and any $(\sigma_1, \ldots, \sigma_n) \in \{-1,1\}^n$,*

$$\mathrm{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{z}_i) - f(\mathbf{z}'_i))\right|\right] \le 2\,\mathrm{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(\mathbf{z}_i)\right|\right].$$

Proof. Let $(\mathbf{z}^a_1, \mathbf{z}^b_1, \ldots, \mathbf{z}^a_n, \mathbf{z}^b_n) \sim P^{2n}$ be a size-$2n$ iid sample from P. We determine $\mathbf{z}_{1:n}$ and $\mathbf{z}'_{1:n}$ as follows. For each $i \in [n]$:

- If $\sigma_i = +1$, then $\mathbf{z}_i := \mathbf{z}^a_i$ and $\mathbf{z}'_i := \mathbf{z}^b_i$.
- If $\sigma_i = -1$, then $\mathbf{z}_i := \mathbf{z}^b_i$ and $\mathbf{z}'_i := \mathbf{z}^a_i$.

Then $\mathbf{z}_{1:n}$ and $\mathbf{z}'_{1:n}$ are independent and each is a size-n iid sample from P. Moreover,

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{z}_i) - f(\mathbf{z}'_i))\right| = \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i(f(\mathbf{z}^a_i) - f(\mathbf{z}^b_i))\right|. \tag{5.4}$$

By the triangle inequality,

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i(f(\mathbf{z}^a_i) - f(\mathbf{z}^b_i))\right| \le \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(\mathbf{z}^a_i)\right| + \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(\mathbf{z}^b_i)\right|. \tag{5.5}$$

Therefore, combining (5.4) and (5.5) and taking expectations, we get

$$\begin{aligned}\mathrm{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{z}_i) - f(\mathbf{z}'_i))\right|\right] &\le \mathrm{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(\mathbf{z}^a_i)\right| + \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(\mathbf{z}^b_i)\right|\right] \\ &= 2\,\mathrm{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(\mathbf{z}_i)\right|\right]\end{aligned}$$

where the last step uses linearity of expectation and the fact that $(\mathbf{z}_1, \ldots, \mathbf{z}_n)$, $(\mathbf{z}^a_1, \ldots, \mathbf{z}^a_n)$, and $(\mathbf{z}^b_1, \ldots, \mathbf{z}^b_n)$ have the same distribution. □

At face value, it may seem that the only purpose of Proposition 5.6 is to bound a quantity involving two size-n iid samples (or a size-$2n$ iid sample) with a quantity involving only one

size-n iid sample. This is helpful, but not the main point. In fact, for any fixed assignment $(\sigma_1, \ldots, \sigma_n) \in \{-1, 1\}^n$, Proposition 5.6 is a "de-symmetrization", as it moves from the symmetric random variable $f(\mathbf{z}_i) - f(\mathbf{z}'_i)$ to the (possibly) asymmetric random variable $\sigma_i f(\mathbf{z}_i)$!

The key is that the bound in Proposition 5.6 holds for every choice of the assignment $(\sigma_1, \ldots, \sigma_n)$, so the right-hand side (without the factor of 2) can be replaced by

$$\frac{1}{2^n} \sum_{\sigma \in \{-1,1\}^n} \mathrm{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(\mathbf{z}_i) \right| \right].$$

If we introduce a random sign vector $(\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_n)$ that is (i) distributed uniformly over $\{-1, 1\}^n$, and (ii) regarded as independent of $\mathbf{z}_{1:n}$, then the above expression is identical to

$$\mathrm{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \boldsymbol{\sigma}_i f(\mathbf{z}_i) \right| \right].$$

The $\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_n$ are called iid *Rademacher random variables*. Combining Propositions 5.5 and 5.6 immediately gives the following proposition.

Proposition 5.7 (Symmetrization by random signs)**.** *Let $\mathbf{z}_{1:n}$ a size-n iid sample from a probability distribution P over $\mathcal{Z}$. Let $(\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_n)$ be n iid Rademacher random variables, independent of $\mathbf{z}_{1:n}$. For any $\mathcal{F} \subseteq \{0, 1\}^{\mathcal{Z}}$,*

$$\mathrm{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(\mathbf{z}_i) - \mathrm{E}[f(\mathbf{z})] \right| \right] \leq 2\,\mathrm{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \boldsymbol{\sigma}_i f(\mathbf{z}_i) \right| \right].$$

The last step is to use the tower property of conditional expectations:

$$\mathrm{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \boldsymbol{\sigma}_i f(\mathbf{z}_i) \right| \right] = \mathrm{E}\left[\mathrm{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \boldsymbol{\sigma}_i f(\mathbf{z}_i) \right| \,\middle|\, \mathbf{z}_{1:n} \right]\right], \tag{5.6}$$

and now focus on the inner conditional expectation. In this conditional expectation, the iid sample $\mathbf{z}_{1:n}$ is fixed. We now (finally!) regard the supremum over $\mathcal{F}$ as a maximum over the behaviors $\mathcal{F}[\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}] \subseteq \{0, 1\}^n$, and we apply the following lemma.

Lemma 5.3 (Massart's finite class lemma)**.** *Let $(\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_n)$ be n iid Rademacher random variables. For any finite subset $V \subseteq \mathbb{R}^n$ with $\|v\| \leq \sqrt{n}$ for all $v \in V$,*

$$\mathrm{E}\left[\max_{(v_1,\ldots,v_n) \in V} \left| \frac{1}{n} \sum_{i=1}^n \boldsymbol{\sigma}_i v_i \right| \right] \leq \sqrt{\frac{2 \ln(2|V|)}{n}}.$$

Proof. For each $v = (v_1, \ldots, v_n) \in V' := V \cup (-V)$, let $\mathbf{x}_v := \frac{1}{n} \sum_{i=1}^n \boldsymbol{\sigma}_i v_i$. Then

$$\max_{(v_1,\ldots,v_n) \in V} \left| \frac{1}{n} \sum_{i=1}^n \boldsymbol{\sigma}_i v_i \right| = \max_{v \in V'} \mathbf{x}_v.$$

For each $v = (v_1, \ldots, v_n) \in V'$, the random variables $(v_1/n)\boldsymbol{\sigma}_1, \ldots, (v_n/n)\boldsymbol{\sigma}_n$ are independent, and $(v_i/n)\boldsymbol{\sigma}_i$ is $(v_i/n)^2$-subgaussian (by Lemma 4.1), so $\mathbf{x}_v$ is $\|v\|^2/n^2$-subgaussian:

$$\mathsf{K}_{\mathbf{x}_v}(\lambda) \leq \frac{\|v\|^2 \lambda^2}{2n^2} \leq \frac{\lambda^2}{2n}.$$

Fix $\lambda > 0$. Then by Jensen's inequality,

$$\begin{aligned}
\mathrm{E}\left[\max_{v\in V'} \mathbf{x}_v\right] &\le \frac{1}{\lambda} \ln \mathrm{E}\left[\exp\left(\lambda \max_{v\in V'} \mathbf{x}_v\right)\right] \\
&= \frac{1}{\lambda} \ln \mathrm{E}\left[\max_{v\in V'} \exp(\lambda \mathbf{x}_v)\right] \\
&\le \frac{1}{\lambda} \ln \mathrm{E}\left[\sum_{v\in V'} \exp(\lambda \mathbf{x}_v)\right] \\
&= \frac{1}{\lambda} \ln\left(\sum_{v\in V'} \exp(\mathsf{K}_{\mathbf{x}_v}(\lambda))\right).
\end{aligned}$$

Now applying the bound on $\mathsf{K}_{\mathbf{x}_v}(\lambda)$ for each $v \in V'$, we obtain

$$\mathrm{E}\left[\max_{v\in V'} \mathbf{x}_v\right] \le \frac{\ln|V'|}{\lambda} + \frac{1}{\lambda} \cdot \frac{\lambda^2}{2n}. = \frac{\ln|V'|}{\lambda} + \frac{\lambda}{2n}.$$

Plugging in $\lambda = \sqrt{2n \ln|V'|}$ and using the fact $|V'| \le 2|V|$ gives the claim. □

Lemma 5.3 gives the following bound on the conditional expectation from (5.6):

$$\mathrm{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\sigma}_i f(\mathbf{z}_i)\right| \,\middle|\, \mathbf{z}_{1:n}\right] \le \sqrt{\frac{2\ln(2|\mathcal{F}[\{\mathbf{z}_1,\ldots,\mathbf{z}_n\}]|)}{n}}.$$

Now we appeal to Lemma 5.2 to bound $|\mathcal{F}[\{\mathbf{z}_1,\ldots,\mathbf{z}_n\}]|$ by $\binom{n}{\le d}$, where $d = \mathsf{VCdim}(\mathcal{F})$. Putting this together with Proposition 5.7 proves Theorem 5.1.

Exercise 5.7. In the proof of Proposition 5.6, if we had taken expectations after (5.4), and then replaced the assignment $(\sigma_1,\ldots,\sigma_n)$ with a random assignment $(\boldsymbol{\sigma}_1,\ldots,\boldsymbol{\sigma}_n)$ (as in Proposition 5.7), we would have gotten the following variation on Proposition 5.7:

$$\mathrm{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(\mathbf{z}_i) - \mathrm{E}[f(\mathbf{z})]\right|\right] = \mathrm{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\sigma}_i(f(\mathbf{z}_i^a) - f(\mathbf{z}_i^b))\right|\right].$$

(a) Explain how to use Massart's finite class lemma (Lemma 5.3) and Sauer's lemma (Lemma 5.2) directly with the right-hand side of the above expression, without passing through Proposition 5.7. How does the result compare to Theorem 5.1?

(b) Show that

$$\begin{aligned}
\mathrm{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\sigma}_i(f(\mathbf{z}_i^a) - f(\mathbf{z}_i^b))\right|\right] &\ge \mathrm{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\sigma}_i(f(\mathbf{z}_i) - \mathrm{E}[f(\mathbf{z})])\right|\right] \\
&\ge \mathrm{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\sigma}_i f(\mathbf{z}_i)\right|\right] - \frac{\sup_{f\in\mathcal{F}}|\mathrm{E}[f(\mathbf{z})]|}{\sqrt{n}}.
\end{aligned}$$ ◊

5.2.5 Zero-one loss class

How can Theorem 5.1 be used to understand the convergence of training error rates to population error rates for hypotheses in a hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$? The key is simply to define a function class $\mathcal{F}_{\mathcal{H}} := \{f_h \mid h \in \mathcal{H}\}$ over the domain $\mathcal{Z} := \mathcal{X} \times \{0,1\}$, where $f_h(x,y) = \mathbb{1}\{h(x) \neq y\}$. This function class $\mathcal{F}_{\mathcal{H}}$ is the *zero-one loss class* associated with $\mathcal{H}$. If P is a probability distribution on $\mathcal{Z} = \mathcal{X} \times \{0,1\}$ and we consider an iid sample from P, then the population averages of $f_h \in \mathcal{F}_{\mathcal{H}}$ is the population error rate of h, and the empirical average of f_h is the training error rate of h.

Proposition 5.8. $\mathsf{VCdim}(\mathcal{F}_{\mathcal{H}}) = \mathsf{VCdim}(\mathcal{H})$.

Proof. We first show that $\mathsf{VCdim}(\mathcal{F}_{\mathcal{H}}) \geq \mathsf{VCdim}(\mathcal{H})$. Take any $T := \{x_1, \ldots, x_d\} \subseteq \mathcal{X}$ shattered by $\mathcal{H}$, i.e., $\mathcal{H}[T] = \{0,1\}^d$. Now define $T' := \{(x_1, 0), \ldots, (x_d, 0)\} \subseteq \mathcal{X} \times \{0,1\}$. For each $h \in \mathcal{H}$, we have $f_h(x_i, 0) = \mathbb{1}\{h(x_i) \neq 0\} = h(x_i)$. So $\mathcal{F}_{\mathcal{H}}[T'] = \mathcal{H}[T] = \{0,1\}^d$, and hence T' is shattered by $\mathcal{F}_{\mathcal{H}}$. This implies $\mathsf{VCdim}(\mathcal{F}_{\mathcal{H}}) \geq \mathsf{VCdim}(\mathcal{H})$.

Now we show that $\mathsf{VCdim}(\mathcal{F}_{\mathcal{H}}) \leq \mathsf{VCdim}(\mathcal{H})$. Take any $T' := \{(x_1, y_1), \ldots, (x_d, y_d)\} \subseteq \mathcal{X} \times \{0,1\}$ of d distinct points shattered by $\mathcal{F}_{\mathcal{H}}$. We want to show that $T := \{x_1, \ldots, x_d\}$ is a set of d points shattered by $\mathcal{H}$.

We first show that $x_1, \ldots, x_d$ are distinct points. To see this, suppose for sake of contradiction that we have $x_i = x_j$ for some $i \neq j$. Then it must be that $y_i \neq y_j$. But for any $h \in \mathcal{H}$, we have

$$f_h((x_i, y_i)) = \mathbb{1}\{h(x_i) \neq y_i\} = 1 - \mathbb{1}\{h(x_i) \neq y_j\} = 1 - \mathbb{1}\{h(x_j) \neq y_j\} = 1 - f_h((x_j, y_j)).$$

This means that no $f_h \in \mathcal{F}_{\mathcal{H}}$ can assign value 1 to both (x_i, y_i) and (x_j, y_j), and hence T' is not shattered by $\mathcal{F}_{\mathcal{H}}$, a contradiction. So, $x_1, \ldots, x_d$ are distinct, and $|T| = d$.

Now we show that T is shattered by $\mathcal{H}$. To see this, consider any labeling $(b_1, \ldots, b_d) \in \{0,1\}^d$. Since T is shattered by $\mathcal{F}_{\mathcal{H}}$, there exists $f_h \in \mathcal{F}_{\mathcal{H}}$ such that $f_h(x_i, y_i) = \mathbb{1}\{b_i \neq y_i\}$ for each $i \in [d]$. But for each $i \in [d]$, we have $f_h(x_i, y_i) = \mathbb{1}\{h(x_i) \neq y_i\}$. Therefore $h(x_i) = b_i$ for each $i \in [d]$. So T is shattered by $\mathcal{H}$. This implies $\mathsf{VCdim}(\mathcal{F}_{\mathcal{H}}) \leq \mathsf{VCdim}(\mathcal{H})$. □

5.2.6 Lower bound for uniform convergence

Exercises 5.6 and 5.7 shows that Proposition 5.7 is tight up to constant factors and an additive $O(1/\sqrt{n})$ term. But what of the bound obtained via Massart's finite class lemma (Lemma 5.3)? This turns out to be almost tight—up to roughly a $\log(n)$ factor—for a worst-case choice of P.

Proposition 5.9. *For any $\mathcal{F} \subseteq \{0,1\}^{\mathcal{Z}}$, there is a probability distribution P over $\mathcal{Z}$ such that, if $\mathbf{z}_{1:n}$ is a size-n iid sample from P, and $(\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_n)$ are n iid Rademacher random variables regarded as independent of $\mathbf{z}_{1:n}$, then*

$$\mathrm{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\sigma}_i f(\mathbf{z}_i) \right| \right] \geq \frac{1}{4} \min\left\{1, \sqrt{\frac{\mathsf{VCdim}(\mathcal{F})}{n}}\right\}.$$

Proof. Let $d := \min\{\mathsf{VCdim}(\mathcal{F}), n\}$. Let $T := \{t_1, \ldots, t_d\}$ be a set of d points shattered by $\mathcal{F}$, and let P be the uniform distribution on T. Define $\mathbf{S}_j := \{i \in [n] \mid \mathbf{z}_i = t_j\}$ and $\mathbf{N}_j := |\mathbf{S}_j|$ for each $j \in [d]$. Then for each $f \in \mathcal{F}$,

$$\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\sigma}_i f(\mathbf{z}_i) = \frac{1}{n} \sum_{j=1}^{d} \sum_{i \in \mathbf{S}_j} \boldsymbol{\sigma}_i f(t_j).$$

Since T is shattered by $\mathcal{F}$, there exists $\mathbf{f}_{\mathrm{o}} \in \mathcal{F}$ such that $\mathbf{f}_{\mathrm{o}}(t_j) = \mathbb{1}\{\sum_{i \in \mathbf{S}_j} \boldsymbol{\sigma}_i > 0\}$ for each $j \in [d]$. Therefore

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\sigma}_i f(\mathbf{z}_i) \right| \geq \left| \frac{1}{n} \sum_{j=1}^{d} \sum_{i \in \mathbf{S}_j} \boldsymbol{\sigma}_i \mathbf{f}_{\mathrm{o}}(t_j) \right| = \frac{1}{n} \sum_{j=1}^{d} \max\left\{ 0, \sum_{i \in \mathbf{S}_j} \boldsymbol{\sigma}_i \right\},$$

and taking expectations,

$$\mathrm{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\sigma}_i f(\mathbf{z}_i) \right| \right] \geq \frac{1}{n} \sum_{j=1}^{d} \mathrm{E}\left[\max\left\{ 0, \sum_{i \in \mathbf{S}_j} \boldsymbol{\sigma}_i \right\} \right].$$

By the symmetry of the distribution of the Rademacher random variables, we have

$$\mathrm{E}\left[\max\left\{ 0, \sum_{i \in \mathbf{S}_j} \boldsymbol{\sigma}_i \right\} \,\middle|\, \mathbf{z}_{1:n} \right] = \frac{1}{2} \mathrm{E}\left[\left| \sum_{i \in \mathbf{S}_j} \boldsymbol{\sigma}_i \right| \,\middle|\, \mathbf{z}_{1:n} \right]$$

A tedious calculation (or Khintchine's inequality) shows that

$$\mathrm{E}\left[\left| \sum_{i \in \mathbf{S}_j} \boldsymbol{\sigma}_i \right| \,\middle|\, \mathbf{z}_{1:n} \right] = \mathbf{N}_j \binom{\mathbf{N}_j - 1}{\lfloor (\mathbf{N}_j - 1)/2 \rfloor} 2^{1 - \mathbf{N}_j} \geq \sqrt{\frac{\mathbf{N}_j}{2}}.$$

Therefore

$$\mathrm{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\sigma}_i f(\mathbf{z}_i) \right| \right] \geq \frac{1}{2\sqrt{2}n} \mathrm{E}\left[\sum_{j=1}^{d} \sqrt{\mathbf{N}_j} \right].$$

By Hölder's inequality, we have

$$n = \sum_{j=1}^{d} \mathbf{N}_j = \sum_{j=1}^{d} \mathbf{N}_j^{1/3} \mathbf{N}_j^{2/3} \leq \left(\sum_{j=1}^{d} \mathbf{N}_j^{1/2} \right)^{2/3} \left(\sum_{j=1}^{d} \mathbf{N}_j^{2} \right)^{1/3},$$

and hence

$$\sum_{j=1}^{d} \sqrt{\mathbf{N}_j} \geq \frac{n^{3/2}}{\sqrt{\sum_{j=1}^{d} \mathbf{N}_j^2}}.$$

By Jensen's inequality,

$$\mathrm{E}\left[\frac{n^{3/2}}{\sqrt{\sum_{j=1}^{d} \mathbf{N}_j^2}} \right] \geq \frac{n^{3/2}}{\sqrt{\mathrm{E}\left[\sum_{j=1}^{d} \mathbf{N}_j^2 \right]}} = \frac{n^{3/2}}{\sqrt{\sum_{j=1}^{d} \mathrm{E}\left[\mathbf{N}_j^2 \right]}}.$$

The random variables $(\mathbf{N}_1, \ldots, \mathbf{N}_d)$ are the occupancies of d bins after throwing n balls independently into uniformly chosen bins, so

$$\mathrm{E}[\mathbf{N}_j^2] = \mathrm{var}(\mathbf{N}_j) + \mathrm{E}[\mathbf{N}_j]^2 = \frac{n}{d}\left(1 - \frac{1}{d}\right) + \left(\frac{n}{d}\right)^2 = \frac{n}{d} + \frac{n(n-1)}{d^2}.$$

We conclude

$$\mathrm{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\sigma}_i f(\mathbf{z}_i) \right| \right] \geq \frac{1}{2\sqrt{2}n} \cdot \frac{n^{3/2}}{\sqrt{n + n(n-1)/d}} = \frac{1}{2\sqrt{2}} \sqrt{\frac{d}{d + n - 1}} \geq \frac{1}{4} \sqrt{\frac{d}{n}}. \qquad \square$$

It shows that $\mathcal{N}(\epsilon, \mathcal{F}, L^\infty(P)) \leq \binom{n}{\leq d}$ even as $\epsilon \to 0$.