Advanced Theory

Uri Shaham

May 31, 2026

1 From Binary Classification to General Function Classes: Rademacher Complexity

The generalization bounds derived from VC dimension are distribution-independent. Although this property universally guarantees their validity across all data distributions, it frequently yields bounds that lack tightness for benign, real-world distributions. Furthermore, the foundational concepts of VC dimension restrict its use to binary classification. To accommodate more complex tasks like multi-class classification and regression, Rademacher complexity provides a modern, distribution-dependent alternative capable of evaluating any class of real-valued functions.

As before, let $\mathcal{D}$ be some distribution on X and let $S = \{x_1, \ldots, x_n\}$ be a set sampled iid from $\mathcal{D}$. Let $\mathcal{F}$ be a function class of functions $f : \mathcal{X} \to \mathbb{R}$.

Definition 1.1 (Empirical Rademacher Complexity)**.** *The empirical Rademacher complexity of $\mathcal{F}$ is defined to be*

$$\hat{R}_n(\mathcal{F}) = \mathbb{E}_{\sigma_1,\ldots\sigma_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{n} \sigma_i f(x_i)\right],$$

where $\sigma_1, \ldots \sigma_n$ are iid random variables distributed unif$(\{-1, 1]\})$ (called Rademacher variables).

The supremum can be viewed as the maximum possible correlation between an element $f \in \mathcal{F}$, viewed as a vector $(f(x_1), \ldots, f(x_n))$ and an instance of the Rademacher vector. Taking the expectation over $\sigma_1, \ldots \sigma_n$ intuitively measures the ability of $\mathcal{F}$ to fit random noise.

Definition 1.2 (Rademacher Complexity)**.** *Rademacher complexity of $\mathcal{F}$ is defined to be*

$$R_n(\mathcal{F}) = \mathbb{E}_{S\sim\mathcal{D}^n}\left[\hat{R}_n(\mathcal{F})\right].$$

The expectation over $S \sim \mathcal{D}^n$ captures the noise-fitting ability of $\mathcal{F}$ over random sets S.

Remark 1.3. *It will sometimes be convenient to talk about Rademacher complexity of sets $A \subset \mathbb{R}^n$ via the supremum $a \in \mathcal{A}$. This generalizes the definition above, by taking $A = \mathcal{F}(S) := \{f(x) : x \in S, f \in \mathcal{F}\}$.*

2 Rademacher-based uniform convergence

We start by mentioning of a concentration bound, which is of interest in its own right.

Theorem 2.1 (McDiarmid inequality)**.** *Let V be some set and let $f : V^n \to \mathbb{R}$ be a function with the property that for some $c > 0$, for all $i \in [n]$ and for all $x_1, \ldots x_n, x_i' \in V$ we have*

$$|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c.$$

Let $X_1, \ldots, X_n \in V$ be independent random variables. Then, with probability at least $1 - \delta$ we have

$$|f(x_1, \ldots x_n) - \mathbb{E}[f(x_1, \ldots x_n)] \leq c\sqrt{\ln\left(\frac{2}{\delta}\right) n/2}.$$

Uniform convergence via Rademacher. Rademacher complexity can let us obtain a uniform convergence result for any class of bounded real-valued functions.

Theorem 2.2. *Assume that for all x and $h \in \mathcal{H}$ we have that $|\ell(h, z)| \leq c$. Then*

1. *with probability at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^m$,*

$$\mathbb{E}_{\mathcal{D}}[h] - \mathbb{E}_S[h] \leq 2R_n(\mathcal{H}) + c\sqrt{\ln\left(\frac{2/\delta}{2n}\right)}.$$

2. *with probability at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^m$,*

$$\mathbb{E}_{\mathcal{D}}[h] - \mathbb{E}_S[h] \leq 2\hat{R}_n(\mathcal{H}) + 3c\sqrt{\ln\left(\frac{4/\delta}{2n}\right)}.$$

Proof. From the definition of supremum, we clearly have that for any fixed $h \in \mathcal{H}$

$$\mathbb{E}_{\mathcal{D}}[h] - \mathbb{E}_S[h] \leq \sup_{h \in \mathcal{H}} \left(\mathbb{E}_{\mathcal{D}}[h] - \mathbb{E}_S[h]\right).$$

Denote $\phi(S) = \sup_{h \in \mathcal{H}} \left(\mathbb{E}_{\mathcal{D}}[h] - \mathbb{E}_S[h]\right)$, and observe that the function $\phi(S)$ satisfies McDiarmid inequality with constant c/m. Thus, with probability at least $1 - \delta$ we have

$$\mathbb{E}_{\mathcal{D}}[h] - \mathbb{E}_S[h] \leq \mathbb{E}_{S \sim \mathcal{D}^m}\left[\sup_{h \in \mathcal{H}} \left(\mathbb{E}_{\mathcal{D}}[h] - \mathbb{E}_S[h]\right)\right] + c\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

To prove the first item, we need to bound $\mathbb{E}_{S \sim \mathcal{D}^n}\left[\sup_{h \in \mathcal{H}} \left(\mathbb{E}_{\mathcal{D}}[h] - \mathbb{E}_S[h]\right)\right]$ in terms of $\mathbb{R}_n(\mathcal{H})$. We do so as follows. Let $S' = \{x_1', \ldots x_n'\}$ be another iid sample. Clearly, for all $h \in \mathcal{H}$ we have $L_{\mathcal{D}}(h) = \mathbb{E}_{S'}[L_{S'}(h)]$, so

$$L_{\mathcal{D}}(h) - L_S(h) = \mathbb{E}_{S'}[L_{S'}(h) - L_S(h)].$$

Next, we use Jensen's inequality, noticing that sup is a convex function, to switch the order of expectation and supremum:

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

Taking expectations over S now, we obtain:

$$\mathbb{E}_S\left[\sup_{h \in \mathcal{H}}(L_{\mathcal{D}}(h) - L_S(h))\right] \leq \mathbb{E}_{S,S'}\left[\sup_{h \in \mathcal{H}}(L_{S'}(h) - L_S(h))\right], \tag{1}$$

and obtain that we can write the RHS as:

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

Note that for each i, z_i and z_i' are iid samples, hence we can switch them without changing the expectation. This means that we can incorporate Rademacher variables into the above expression and write it as

$$\begin{aligned}
\frac{1}{n}\,\mathbb{E}_{S,S',\sigma}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{n}\sigma_i(\ell(h,z_i')-\ell(h,z_i))\right] &\le \frac{1}{m}\,\mathbb{E}_{S,S',\sigma}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{n}\sigma_i\ell(h,z_i')-\sup_{h\in\mathcal{H}}\sum_{i=1}^{n}\sigma_i\ell(h,z_i)\right]\\
&= \frac{1}{n}\,\mathbb{E}_{S,\sigma}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{n}\sigma_i\ell(h,z_i')\right]+\frac{1}{m}\,\mathbb{E}_{S',\sigma}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{n}-\sigma_i\ell(h,z_i)\right]\\
&= \frac{1}{n}\,\mathbb{E}_{S,\sigma}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{n}\sigma_i\ell(h,z_i')\right]+\frac{1}{n}\,\mathbb{E}_{S',\sigma}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{n}\sigma_i\ell(h,z_i)\right]\\
&= 2R_n(\mathcal{H}).
\end{aligned}$$

Substituting this into equation (2) proves the first statement of the theorem.

To prove the second statement, observe that $\hat{R}_m(\mathcal{H})$ satisfies the McDiarmid inequality (this time with confidence $\delta/2$) with constant c/n. Thus, using McDiarmid to replace each $R_m(\mathcal{H})$ by $\hat{R}_m(\mathcal{H})$ and then the union bound (sum of the two probabilities) results in the second statement. □

Growth of the Rademacher Complexity - Massart Lemma. Massart's lemma shows that the Rademacher complexity of a finite set grows logarithmically with the size of the set.

Lemma 2.3 (Massart lemma)**.** *Let $A\subseteq\mathbb{R}^m$ be finite. Let $R=\max_{a\in A}\|a\|_2$. Then*

$$\hat{R}_n(A)=\mathbb{E}_\sigma\left[\sup_{a\in A}\left(\frac{1}{m}\sigma_i a_i\right)\right]\le R\frac{\sqrt{2\ln|A|}}{n}.$$

To prove the Massart lemma, we will make use of Hoeffding's lemma:

Lemma 2.4 (Hoeffding's lemma)**.** *Let X be a bounded random variable such that $a\le X\le b$. Then for all $\lambda\in\mathbb{R}$,*

$$\mathbb{E}\left[e^{\lambda X}\right]\le\exp\left(\lambda\,\mathbb{E}[X]+\frac{\lambda^2(b-a)^2}{8}\right).$$

We can now prove Massart lemma:

Proof. We start by taking the exponential of the Rademacher complexity times some positive constant s that we will choose later:

$$
\begin{aligned}
\exp\left(s\,\mathbb{E}_\sigma\left[\sup_{a\in A}\sum_{i=1}^{n}\sigma_i a_i\right]\right) &\le \mathbb{E}_\sigma\left[\exp\left(s\sup_{a\in A}\sum_{i=1}^{n}\sigma_i a_i\right)\right] \text{ (Jensen)}\\
&= \mathbb{E}_\sigma\left[\sup_{a\in A}\exp\left(s\sum_{i=1}^{n}\sigma_i a_i\right)\right]\\
&\le \sum_{a\in A}\mathbb{E}_\sigma\left[\exp\left(s\sum_{i=1}^{n}\sigma_i a_i\right)\right]\\
&= \sum_{a\in A}\mathbb{E}_\sigma\left[\prod_{i=1}^{n}e^{(s\sigma_i a_i)}\right]\\
&= \sum_{a\in A}\prod_{i=1}^{n}\mathbb{E}_{\sigma_i}\left[e^{s\sigma_i a_i}\right] \text{ (Independence of the } \sigma_i\text{'s)}\\
&\le \sum_{a\in A}\prod_{i=1}^{n}\exp\left(\frac{4s^2a_i^2}{8}\right) \text{ (Hoeffding's lemma, } b-a=2sa_i)\\
&= \sum_{a\in A}\exp\left(\frac{s^2}{2}\sum_{i=1}^{n}a_i^2\right)\\
&= |A|\exp\left(\frac{s^2R^2}{2}\right).
\end{aligned}
$$

Now taking log of both sides and dividing by s gives

$$
\mathbb{E}_\sigma\left[\sup_{a\in A}\sum_{i=1}^{m}\sigma_i a_i\right] \le \frac{\ln(|A|)}{s}+\frac{sR^2}{2}.
$$

We now take the derivative of the right-hand side with respect to s and equate it to zero, which gives $s=\frac{\sqrt{2\ln(|A|}}{R}$, and substitute this back to get

$$
\mathbb{E}_\sigma\left[\sup_{a\in A}\sum_{i=1}^{n}\sigma_i a_i\right] \le R\sqrt{2\ln(|A|}.
$$

Dividing both sides by n concludes the proof.

□

The main takeaway from the Massart lemma is that the generalization error of a finite hypothesis class $\mathcal{H}$ scales like $\frac{\ln(\mathcal{H})}{m}$, i.e., the training set size m needs to grow like $\ln(\mathcal{H})$.

3 From Finite to Infinite Classes: Covering numbers

3.1 Motivation

The strategy is straightforward: *approximate* an infinite class by a finite one, and pay a small price for the approximation error.

Fix a sample S and suppose we find a finite set $\mathcal{H}_\varepsilon \subset \mathcal{H}$ such that for every $h \in \mathcal{H}$ there exists some $h' \in \mathcal{H}_\varepsilon$ with

$$\frac{1}{n}\sum_{i=1}^{n}\big(h(x_i) - h'(x_i)\big)^2 \le \varepsilon^2.$$

Then, roughly, the supremum over $\mathcal{H}$ behaves like the supremum over $\mathcal{H}_\varepsilon$ up to an error of order ε, and Massart applies to $\mathcal{H}_\varepsilon$. This is the idea of *covering numbers*.

3.2 Covering Numbers: Definitions

Definition 3.1 (ε-Cover)**.** *Let $(\mathcal{F}, d)$ be a metric space. A set $\mathcal{F}_\varepsilon$ is an* ε-cover *of $\mathcal{F}$ with respect to d if for every $f \in \mathcal{F}$ there exists $f' \in \mathcal{F}_\varepsilon$ such that $d(f, f') \le \varepsilon$.*

The ε-covering number *is*

$$\mathcal{N}(\varepsilon, \mathcal{F}, d) = \min\big\{|\mathcal{F}_\varepsilon| : \mathcal{F}_\varepsilon \text{ is an } \varepsilon\text{-cover of } \mathcal{F}\big\}.$$

The metric entropy *of $\mathcal{F}$ at scale ε is* $\log \mathcal{N}(\varepsilon, \mathcal{F}, d)$.

Definition 3.2 (ε-Packing)**.** *A set $\mathcal{F}_\varepsilon$ is an* ε-packing *of $\mathcal{F}$ if every two distinct $f, f' \in \mathcal{F}_\varepsilon$ satisfy $d(f, f') > \varepsilon$. The* ε-packing number $\mathcal{M}(\varepsilon, \mathcal{F}, d)$ *is the size of the largest such packing.*

Covering and packing numbers are related by the following standard sandwich inequality, which means we can use whichever is more convenient.

Proposition 3.3 (Covering-Packing Equivalence)**.** *For any $\varepsilon > 0$,*

$$\mathcal{M}(2\varepsilon, \mathcal{F}, d) \le \mathcal{N}(\varepsilon, \mathcal{F}, d) \le \mathcal{M}(\varepsilon, \mathcal{F}, d).$$

Proof. **Left inequality.** Let P be a maximal 2ε-packing. Each ball of radius 2ε around a packing point contains at most one other packing point. If the cover C had $|C| < |P|$, by pigeonhole some cover point c would be within ε of two packing points p, p', giving $d(p, p') \le 2\varepsilon$, a contradiction. So $|C| \ge |P|$.

Right inequality. A maximal ε-packing is also an ε-cover: if some f were not covered, it would be at distance $> \varepsilon$ from all packing points, contradicting maximality. □

In statistical learning theory, the relevant metric on a sample $S = (x_1, \dots, x_n)$ is the empirical L_2 metric:

$$d_S(h, h') = \left(\frac{1}{n}\sum_{i=1}^{n}\big(h(x_i) - h'(x_i)\big)^2\right)^{1/2}.$$

We write $\mathcal{N}(\varepsilon, \mathcal{H}, d_S)$ for the covering number of $\mathcal{H}$ under this metric.

3.3 Single-Scale Bound

We now derive the simplest form of the covering-number bound on Rademacher complexity. This already captures the main idea and is a direct application of Massart's lemma.

Theorem 3.4 (Single-Scale Bound)**.** *Let $\mathcal{H}$ be a class of functions $h : \mathcal{X} \to [-1, 1]$ and let $S = (x_1, \dots, x_n)$ be a fixed sample. Then for any $\varepsilon > 0$,*

$$\hat{\mathcal{R}}_S(\mathcal{H}) \le \varepsilon + \sqrt{\frac{2\log \mathcal{N}(\varepsilon, \mathcal{H}, d_S)}{n}}.$$

Proof. Let $\mathcal{H}_\varepsilon$ be an ε-cover of $\mathcal{H}$ under d_S of size $\mathcal{N}(\varepsilon, \mathcal{H}, d_S)$. For any $h \in \mathcal{H}$, let $\pi(h) \in \mathcal{H}_\varepsilon$ denote a nearest cover point, so that $d_S(h, \pi(h)) \le \varepsilon$.

We bound the Rademacher complexity by decomposing each h as $h = \pi(h) + (h - \pi(h))$:

$$\begin{aligned}
\hat{\mathcal{R}}_S(\mathcal{H}) &= \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \sigma_i h(x_i)\right] \\
&= \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \left(\frac{1}{n}\sum_{i=1}^n \sigma_i \pi(h)(x_i) + \frac{1}{n}\sum_{i=1}^n \sigma_i (h-\pi(h))(x_i)\right)\right] \\
&\le \mathbb{E}_\sigma\left[\sup_{h'\in\mathcal{H}_\varepsilon} \frac{1}{n}\sum_{i=1}^n \sigma_i h'(x_i)\right] + \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\left|\sum_{i=1}^n \sigma_i (h-\pi(h))(x_i)\right|\right].
\end{aligned}$$

First term. Apply Massart's lemma to the finite class $\mathcal{H}_\varepsilon$, with each vector $a_{h'} = \frac{1}{n}(h'(x_1), \dots, h'(x_n)) \in \mathbb{R}^n$ satisfying $\|a_{h'}\|_2 \le \frac{1}{\sqrt{n}}$:

$$\mathbb{E}_\sigma\left[\sup_{h'\in\mathcal{H}_\varepsilon} \frac{1}{n}\sum_{i=1}^n \sigma_i h'(x_i)\right] \le \sqrt{\frac{2\log|\mathcal{H}_\varepsilon|}{n}} = \sqrt{\frac{2\log\mathcal{N}(\varepsilon,\mathcal{H},d_S)}{n}}.$$

Second term. By the Cauchy–Schwarz inequality, for any h:

$$\frac{1}{n}\left|\sum_{i=1}^n \sigma_i (h-\pi(h))(x_i)\right| \le \frac{1}{n}\|\sigma\|_2 \cdot \|(h-\pi(h))|_S\|_2 = d_S(h,\pi(h)) \le \varepsilon,$$

where we used $\|\sigma\|_2 = \sqrt{n}$ and the definition of d_S. Hence this term contributes at most ε.

Combining gives the result. □

Remark 3.5. *The bound involves a bias-variance-like tradeoff: larger ε means a coarser cover (smaller $\log\mathcal{N}$) but larger approximation error. We can optimize over ε to get the best bound for a given class, as we will see in examples.*

3.4 Dudley's Entropy Integral

The single-scale bound is already useful, but it is suboptimal. It captures the class at only one resolution ε. A finer idea — *chaining* — accounts for the geometry of $\mathcal{H}$ at all scales simultaneously.

3.5 Intuition: Chaining

Instead of approximating each h by a single cover point at scale ε, we build a sequence of increasingly refined covers $\mathcal{H}_{\varepsilon_0} \supset \mathcal{H}_{\varepsilon_1} \supset \cdots$ with $\varepsilon_0 > \varepsilon_1 > \cdots \to 0$. We write each h as a telescoping sum:

$$h = h_0 + (h_1 - h_0) + (h_2 - h_1) + \cdots,$$

where $h_k = \pi_{\varepsilon_k}(h)$ is the projection of h onto the cover at scale ε_k. Each increment $h_k - h_{k-1}$ is small (of size $\sim \varepsilon_{k-1}$), and the number of possible increments at level k is at most $\mathcal{N}(\varepsilon_k, \mathcal{H}, d_S)$. Summing the Massart bounds over levels yields an integral of the metric entropy.

3.6 The Theorem

Theorem 3.6 (Dudley's Entropy Integral)**.** *Let $\mathcal{H}$ be a class of functions $h : \mathcal{X} \to [-1, 1]$. Then*

$$\hat{\mathcal{R}}_S(\mathcal{H}) \leq \inf_{\delta \geq 0} \left(4\delta + \frac{12}{\sqrt{n}} \int_\delta^1 \sqrt{\log \mathcal{N}(\varepsilon, \mathcal{H}, d_S)}\, d\varepsilon \right).$$

Proof sketch. Set $\varepsilon_k = 2^{-k}$ and let $\mathcal{H}_k$ be an ε_k-cover of $\mathcal{H}$, with $|\mathcal{H}_k| = \mathcal{N}(\varepsilon_k, \mathcal{H}, d_S)$. Let $\pi_k : \mathcal{H} \to \mathcal{H}_k$ denote projection onto the nearest cover point at level k.

Fix some level k_0 corresponding to $\delta = \varepsilon_{k_0}$. For any $h \in \mathcal{H}$, write the telescoping decomposition:

$$h = \pi_{k_0}(h) + \sum_{k > k_0} \big(\pi_k(h) - \pi_{k-1}(h) \big).$$

Note that $d_S(\pi_k(h), \pi_{k-1}(h)) \leq d_S(\pi_k(h), h) + d_S(h, \pi_{k-1}(h)) \leq \varepsilon_k + \varepsilon_{k-1} \leq 3\varepsilon_{k-1}$, so the k-th increment is small.

Taking the supremum and expectation, and applying the triangle inequality:

$$\hat{\mathcal{R}}_S(\mathcal{H}) \leq \hat{\mathcal{R}}_S(\mathcal{H}_{k_0}) + \sum_{k > k_0} \hat{\mathcal{R}}_S\big(\{ \pi_k(h) - \pi_{k-1}(h) : h \in \mathcal{H} \} \big).$$

The first term is bounded by 4δ (since d_S-diameter of $\mathcal{H}_{k_0}$ is at most 2δ). For each subsequent term, the number of possible increments $\pi_k(h) - \pi_{k-1}(h)$ is at most $|\mathcal{H}_k| \cdot |\mathcal{H}_{k-1}| \leq \mathcal{N}(\varepsilon_k, \mathcal{H}, d_S)^2$, and each increment has d_S-norm at most $3\varepsilon_{k-1}$. Massart gives a contribution of at most

$$\frac{3\varepsilon_{k-1}}{\sqrt{n}} \cdot \sqrt{2 \cdot 2 \log \mathcal{N}(\varepsilon_k, \mathcal{H}, d_S)}.$$

Summing over k and bounding the sum by an integral (since $\varepsilon_k = 2^{-k}$ is geometrically spaced) yields:

$$\sum_{k > k_0} (\cdots) \leq \frac{12}{\sqrt{n}} \int_\delta^1 \sqrt{\log \mathcal{N}(\varepsilon, \mathcal{H}, d_S)}\, d\varepsilon.$$

Taking the infimum over δ completes the proof. □

Remark 3.7. *The key structural insight is that the bound is governed by the* integral *of $\sqrt{\text{metric entropy}}$, not just its value at one scale. Classes with rapidly decaying metric entropy (as $\varepsilon \to 0$) have small Rademacher complexity; classes where $\log \mathcal{N}(\varepsilon)$ grows slowly are harder to learn.*

Remark 3.8 (Convergence of the integral)**.** *The integral $\int_0^1 \sqrt{\log \mathcal{N}(\varepsilon, \mathcal{H}, d)}\, d\varepsilon$ need not converge. For instance, if $\log \mathcal{N}(\varepsilon) \sim 1/\varepsilon^2$, the integrand grows as $1/\varepsilon$ and the integral diverges. In such cases the δ-cutoff in the infimum is doing real work: it regularizes the bound by truncating at a scale $\delta > 0$, which then shows up as an additive error 4δ.*

4 Covering Numbers vs. VC Dimension

It is instructive to compare the two major complexity measures for hypothesis classes.

	VC Dimension	**Covering Numbers**
Applies to	Binary classification	General function classes
Bound flavor	Combinatorial	Metric / geometric
Main lemma	Sauer–Shelah	Massart + chaining
Main theorem	VC generalization bound	Dudley entropy integral
Granularity	Single number d	Full scale profile $\mathcal{N}(\varepsilon, \mathcal{H}, d)$
Captures geometry?	No	Yes

Covering numbers are strictly more expressive. The VC dimension gives a single combinatorial number that is the same regardless of the metric structure of $\mathcal{H}$, whereas $\mathcal{N}(\varepsilon, \mathcal{H}, d)$ describes how $\mathcal{H}$ looks at every resolution. For function-valued classes (regression, neural networks, Lipschitz families), VC dimension does not apply at all, while covering numbers extend naturally.

On the other hand, VC dimension is often easier to compute from the structure of a class, while covering numbers may require more geometric work.

Appendix: Computing Covering Numbers

We now work through three concrete examples showing how to compute $\mathcal{N}(\varepsilon, \mathcal{H}, \cdot)$ and what bounds result.

The Euclidean Ball

Example 4.1 (Covering the ℓ_2 ball)**.** *Consider* $B_2^d = \{w \in \mathbb{R}^d : \|w\|_2 \leq 1\}$.

Proposition 4.2. $\mathcal{N}(\varepsilon, B_2^d, \|\cdot\|_2) \leq \left(\frac{3}{\varepsilon}\right)^d$.

Proof. Consider a maximal ε-packing $\{w_1, \ldots, w_N\}$ of B_2^d. The balls $B(w_i, \varepsilon/2)$ are pairwise disjoint and all lie in $B_2^d(1 + \varepsilon/2) \subseteq B_2^d(3/2)$ (since $\|w_i\|_2 \leq 1$ and the ball has radius $\varepsilon/2 \leq 1/2$). Comparing volumes:

$$N \cdot \mathrm{vol}(B(\varepsilon/2)) \leq \mathrm{vol}(B(3/2)) \implies N \leq \left(\frac{3/2}{\varepsilon/2}\right)^d = \left(\frac{3}{\varepsilon}\right)^d.$$

Since the packing number upper bounds the covering number (Proposition 1), we are done. □

The metric entropy is thus $\log \mathcal{N}(\varepsilon) \leq d \log(3/\varepsilon)$.

Lipschitz Functions on $[0, 1]$

Example 4.3 (Lipschitz functions)**.** *Let* $\mathcal{F}_L = \{f : [0,1] \to [-1,1] : |f(x) - f(y)| \leq L|x - y|\}$.

Proposition 4.4. $\mathcal{N}(\varepsilon, \mathcal{F}_L, \|\cdot\|_\infty) \leq e^{O(L/\varepsilon)}$.

Proof idea. Partition $[0,1]$ into $m = \lceil L/\varepsilon \rceil$ equally spaced intervals. On each interval, an L-Lipschitz function varies by at most $L/m \le \varepsilon$. Discretizing the value at each grid point to a grid of spacing ε gives an ε-cover of size at most $(2/\varepsilon + 1)^m = e^{O(m \log(1/\varepsilon))} = e^{O(L/\varepsilon \cdot \log(1/\varepsilon))}$. A more careful argument removes the extra log factor. □

Here $\log \mathcal{N}(\varepsilon) = O(L/\varepsilon)$, so $\sqrt{\log \mathcal{N}(\varepsilon)} = O(\sqrt{L/\varepsilon})$. Attempting to apply Dudley:

$$\int_\delta^1 \sqrt{\log \mathcal{N}(\varepsilon)}\, d\varepsilon = O\Big(\sqrt{L} \int_\delta^1 \varepsilon^{-1/2}\, d\varepsilon\Big) = O\Big(\sqrt{L}(1 - \sqrt{\delta})\Big).$$

This is $O(\sqrt{L})$ as $\delta \to 0$, so the integral *converges* and we get $\hat{\mathcal{R}}_S(\mathcal{F}_L) = O(\sqrt{L/n})$

VC Classes via Sauer Lemma

Example 4.5 (VC classes)**.** *Let $\mathcal{H}$ be a binary hypothesis class with VC dimension $VC(\mathcal{H}) = d$.*

The connection between VC dimension and covering numbers runs through the Sauer–Shelah lemma, which bounds the number of distinct labelings a VC-d class can produce.

Lemma 4.6 (Sauer–Shelah)**.** *If $VC(\mathcal{H}) = d$, then for any sample S of size n,*

$$|\{(h(x_1), \dots, h(x_n)) : h \in \mathcal{H}\}| \le \sum_{k=0}^{d} \binom{n}{k} \le \Big(\frac{en}{d}\Big)^d.$$

On a sample of size n, a VC-d class has at most $(en/d)^d$ distinct behaviors, so its covering number satisfies:

$$\mathcal{N}(\varepsilon, \mathcal{H}, d_S) \le \Big(\frac{en}{d}\Big)^d \quad \text{for any } \varepsilon > 0.$$

Massart's lemma then immediately gives:

$$\hat{\mathcal{R}}_S(\mathcal{H}) \le \sqrt{\frac{2d \log(en/d)}{n}} = O\left(\sqrt{\frac{d \log n}{n}}\right),$$

recovering the standard VC generalization bound. Notice that covering numbers provide a unified framework from which VC-type bounds emerge as a special case.

5 Reading

- UML ch. 7
- `https://www.cs.cmu.edu/~ninamf/ML11/lect1117.pdf`