IDS.160 – Mathematical Statistics: A Non-Asymptotic Approach

Lecturer: A Rakhlin
Scribe: A. Rakhlin

Lecture 18 & 19
Apr. 14 & 16, 2020

Goals: We continue the investigation of model complexity through the lens of covering/packing numbers and combinatorial dimensions. These are convenient tools for upper/lower bounding the supremum of the Rademacher or empirical process.

1. COVERING AND PACKING

Given a probability measure P on $\mathcal{X}$, we define

$$\|f\|_{L^2(P)}^2 = \mathbb{E}f(X)^2 = \int f(x)^2 P(dx).$$

Similarly, for a given $X_1, \dots, X_n$ we define a random pseudometric

$$\|f\|_{L^2(P_n)}^2 = \frac{1}{n}\sum_{i=1}^n f(X_i)^2 = \|f\|_n^2 .$$

Of course, the second definition is just a special case of the first for empirical measure $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}$.

Definition: An ε-net (or, ε-cover) of $\mathcal{F}$ with respect to $L^2(P)$ is a set of functions $f_1, \dots, f_N$ such that

$$\forall f \in \mathcal{F}, \quad \exists j \in [N] \quad \text{s.t.} \quad \|f - f_j\|_{L^2(P)} \le \varepsilon.$$

The size of the smallest ε-net is denoted by $\mathcal{N}(\mathcal{F}, L^2(P), \varepsilon)$.

The above definition can be also generalized to $L^p(P)$. Next, we spell out the above definition specifically for the empirical measure P_n:

Definition: Let $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ be the empirical measure supported on $x_1, \dots, x_n$. A set $V = \{v_1, \dots, v_N\}$ of vectors in $\mathbb{R}^n$ forms an ε-net (or, ε-cover) of $\mathcal{F}$ with respect to $L^p(P_n)$ if

$$\forall f \in \mathcal{F}, \quad \exists j \in [N] \quad \text{s.t.} \quad \frac{1}{n}\sum_{i=1}^n |f(x_i) - v_j(i)|^p \le \varepsilon^p$$

The size of the smallest ε-net is denoted by $\mathcal{N}(\mathcal{F}, L^p(P_n), \varepsilon)$. Similarly, an ε-net (or, ε-cover) with respect to $L^\infty(P_n)$ requires

$$\forall f \in \mathcal{F}, \quad \exists j \in [N] \quad \text{s.t.} \quad \max_{i \in [n]} |f(x_i) - v_j(i)| \le \varepsilon$$

The size of the smallest ε-net is denoted by $\mathcal{N}(\mathcal{F}, L^\infty(P_n), \varepsilon)$.

Observe that the elements of the cover V can be "improper," i.e. they do not need to correspond to values of some function on the data. However, one can go between proper and improper covers at a cost of a constant (check!).

Second, observe that

$$\mathcal{N}(\mathcal{F}, L^p(P_n), \varepsilon) \leq \mathcal{N}(\mathcal{F}, L^q(P_n), \varepsilon)$$

for $p \leq q$ since $\|f\|_{L^p(P_n)}$ increases with p. Note that this is different for unweighted metrics: e.g. $\|x\|_p$ is nonincreasing in p, and hence $\mathcal{N}(\Theta, \|\cdot\|_p, \varepsilon)$ is also nonincreasing in p.

Definition: An ε-packing of $\mathcal{F}$ with respect to $L^p(P_n)$ is a set $f_1, \ldots, f_N \in \mathcal{F}$ such that

$$\frac{1}{n}\sum_{i=1}^n |f_j(x_i) - f_k(x_i)|^p \geq \varepsilon^p$$

for any $j \neq k$. The size of the largest ε-packing is denoted by $\mathcal{D}(\mathcal{F}, L^p(P_n), \varepsilon)$.

A standard relationship between covering and packing holds for any P:

$$\mathcal{D}(\mathcal{F}, L^p(P), 2\varepsilon) \leq \mathcal{N}(\mathcal{F}, L^p(P), \varepsilon) \leq \mathcal{D}(\mathcal{F}, L^p(P), \varepsilon)$$

2. UPPER AND LOWER BOUNDS FOR RADEMACHER AVERAGES

As before, we let $U_\theta = \langle \epsilon, \theta \rangle$, $\Theta = \frac{1}{\sqrt{n}}\mathcal{F}|_{x_1,\ldots,x_n}$, and d Euclidean distance. Then from last lecture

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(x_i) &= \mathbb{E}\sup_{\theta\in\Theta} U_\theta \\ &\leq 2\delta\sqrt{n} + 8\sqrt{2}\int_{\delta/4}^{D/2} \sqrt{\log \mathcal{N}(\Theta, d, \varepsilon)}d\varepsilon\end{aligned}$$

Trivially,

$$\mathcal{N}(\Theta, d, \varepsilon) = \mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon).$$

Corollary: For any $X_1, \ldots, X_n$,

$$\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) \leq \inf_{\delta\geq 0}\left\{8\delta + \frac{12}{\sqrt{n}}\int_\delta^{D/2} \sqrt{\log \mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon)}d\varepsilon\right\}$$

with $D = \sup_{f,g\in\mathcal{F}} \|f - g\|_n \leq 2\sup_{f\in\mathcal{F}} \|f\|_n \leq 2\sup_{f\in\mathcal{F}} \|f\|_\infty$.

Putting together the symmetrization lemma and above Corollary, we have

Corollary: Let $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$ be a class of functions and let $X_1, \ldots, X_n \sim P$ be independent. Then

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}f(X) - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right\} \leq \mathbb{E} \inf_{\delta \geq 0} \left\{ 16\delta + \frac{24}{\sqrt{n}} \int_{\delta}^{D} \sqrt{\log \mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon)} d\varepsilon \right\} \tag{2.1}$$

where $D = \sup_{f \in \mathcal{F}} \sqrt{\frac{1}{n} \sum_{i=1}^{n} f(X_i)^2}$.

Expectations on both sides are with respect to $X_1, \ldots, X_n$. Note that the above results hold for the absolute value of the empirical process if we replace $\log \mathcal{N}$ by $\log 2\mathcal{N}$, and the $\log 2$ can be further absorbed into the multiplicative constant.

The Sudakov lower bound for the Gaussian process implies (together with the relationship between Rademacher and Gaussian processes) the following lower bound for the Rademacher averages:

Corollary: For any $X_1, \ldots, X_n$,

$$\mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \geq \frac{c}{\sqrt{\log n}} \cdot \sup_{\alpha \geq 0} \alpha \sqrt{\frac{\log \mathcal{N}(\mathcal{F}, L^2(P_n), \alpha)}{n}}$$

for some absolute constant c.

We note that a version of the lower bound (for a particular choice of α) without the logarithmic factor is available, under some conditions, and it often matches the upper bound (see a few pages below).

3. PARAMETRIC AND NONPARAMETRIC CLASSES OF FUNCTIONS

There is no clear definition of what constitutes a "nonparametric class," especially since the same class of functions (e.g. neural networks) can be treated as either parametric or nonparametric (e.g. if neural network complexity is measured by matrix norms rather than number of parameters).

Consider the following (slightly vague) definition as a possibility:

Definition: We will say that a class $\mathcal{F}$ is *parametric* if for any empirical measure P_n,

$$\mathcal{N}(\mathcal{F}, L^2(P_n), \epsilon) \lesssim \left(\frac{1}{\epsilon}\right)^{\dim}.$$

We will say that $\mathcal{F}$ is *nonparametric* if for any empirical measure P_n,

$$\log \mathcal{N}(\mathcal{F}, L^2(P_n), \epsilon) \asymp \left(\frac{1}{\epsilon}\right)^{p}. \tag{3.2}$$

The requirement that (3.2) holds for all measures P_n and values of n is quite strong. Yet, we will show that as an upper bound, it is true for a variety of function classes. However, one should keep in mind that there are also cases where dependence of the upper bound on n can lead to better overall estimates. The quantity

$$\sup_Q \log \mathcal{N}(\mathcal{F}, L^2(Q), \epsilon),$$

where supremum is taken over all discrete measures, is called *Koltchinskii-Pollard entropy*.

Let's consider a "parametric" class $\mathcal{F}$ such that functions in $\mathcal{F}$ are uniformly bounded: $|f|_\infty \leq 1$. This provides an upper bound on the diameter: $D/2 \leq 1$. Then, taking $\delta = 0$, conditionally on $X_1, \ldots, X_n$,

$$\begin{aligned}
\mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) &\leq \frac{12}{\sqrt{n}} \int_0^1 \sqrt{\log \mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon)} d\varepsilon \\
&\leq \frac{12}{\sqrt{n}} \int_0^1 \sqrt{d \log(1/\epsilon)} d\varepsilon \\
&\leq c \sqrt{\frac{d}{n}}
\end{aligned}$$

Here it's useful to note that

$$\int_0^a \sqrt{\log(1/\varepsilon)} d\varepsilon \leq \begin{cases} 2a\sqrt{\log(1/a)} & a \leq 1/e \\ 2a & a > 1/e \end{cases}$$

The following theorem is due to D. Haussler (an earlier version with exponent $O(d)$ is due to Dudley '78):

Theorem: Let $\mathcal{F} = \{f : \mathcal{X} \to \{0, 1\}\}$ be a class of binary-valued functions with VC dimension $\text{vc}(\mathcal{F}) = d$. Then for any n and any P_n,

$$\mathcal{N}(\mathcal{F}, L^2(P_n), \epsilon) \leq Cd(4e)^d \left(\frac{1}{\epsilon}\right)^{2d}.$$

We will explain what "VC dimension" means a bit later, and let's just say here that the class of thresholds has dimension 1 and the class of homogenous linear classifiers in $\mathbb{R}^d$ has dimension d. In particular, this removes the extraneous $\log(n+1)$ factor we had in Lecture 14 when analyzing thresholds.

3.1 A phase transition

Let us inspect the Dudley integral upper bound. Note that when we plug in

$$\log \mathcal{N}(\mathcal{F}, L^2(P_n), \epsilon) \lesssim \left(\frac{1}{\epsilon}\right)^p,$$

the integral becomes

$$\int_\delta^{D/2} \varepsilon^{-p/2} d\varepsilon$$

If $p < 2$, the integral converges, and we can take $\delta = 0$. However, when $p > 2$, the lower limit of the integral matters and we get an overall bound of the order

$$\delta + n^{-1/2} \left[\varepsilon^{1-p/2}\right]_{\delta}^{D/2} \leq \delta + n^{-1/2}\delta^{1-p/2}$$

By choosing δ to balance the two terms (and thus minimize the upper bound) we obtain $\delta = n^{-1/p}$. Hence, for $p > 2$, the estimate on Rademacher averages provided by the Dudley bound is

$$\widehat{\mathcal{R}}(\mathcal{F}) \lesssim n^{-1/p}.$$

On the other hand, for $p < 2$, the Dudley entropy integral upper bound becomes (by setting $\delta = 0$) on the order of

$$n^{-1/2}D^{1-p/2} = O(n^{-1/2}),$$

yielding

$$\widehat{\mathcal{R}}(\mathcal{F}) \lesssim n^{-1/2}.$$

We see that there is a transition at $p = 2$ in terms of the growth of Rademacher averages ("elbow" behavior). The phase transition will be important in the rest of the course when we study optimality of nonparametric least squares.

Remark that in the $p < 2$ regime, the rate $n^{-1/2}$ is the same rate CLT rate we would have if we simply considered $\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f\right|$ (or the average with random signs) with a single function. Hence, the payment for the supremum over class $\mathcal{F}$ is only in a constant that doesn't depend on n.

3.2 Single scale vs chaining

It is also worthwhile to compare the single-scale upper bound we obtained earlier to the tighter upper bound given by chaining. In other words, we are comparing

$$\delta + \sqrt{\frac{\log \mathcal{N}(\delta)}{n}}$$

versus

$$\delta + \int_{\delta}^{D/2} \sqrt{\frac{\log \mathcal{N}(\varepsilon)}{n}} d\varepsilon,$$

simplifying the notation for brevity.

In the parametric case, the single-scale bound becomes (with the choice of $\delta = 1/n$)

$$\sqrt{\frac{\dim \ \log n}{n}}$$

while chaining gives

$$\sqrt{\frac{\dim}{n}}.$$

In the nonparametric case, the difference is more stark:

$$\delta + \sqrt{\frac{\delta^{-p}}{n}} \asymp n^{-\frac{1}{2+p}}$$

vs

$$n^{-1/2}$$

for $p < 2$, and

$$\delta + \frac{\delta^{1-p/2}}{\sqrt{n}} \asymp n^{-1/p}$$

for $p > 2$.

3.3 Linear class: Parametric or Nonparametric?

Let's take a closer look at the function class

$$\mathcal{F} = \{x \mapsto \langle w, x\rangle : w \in \mathsf{B}_2^d\}$$

and take $\mathcal{X} = \mathsf{B}_2^d$. Recall that for a given $x_1, \ldots, x_n$,

$$\mathcal{F}|_{x_1,\ldots,x_n} = \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\} = \left\{Xw : w \in \mathsf{B}_2^d\right\}$$

where X is the $n \times d$ data matrix. As we have seen, the key quantity we need to compute is

$$\mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon).$$

What is a good upper bound for this quantity? What we had done in Lecture 16 was to discretize the set B_2^d to create a ε-net $w_1, \ldots, w_N$ of size $\mathcal{N}(\mathsf{B}_2^d, \|\cdot\|_2, \varepsilon)$. Clearly, for any w and the corresponding ε-close element w_j of the cover,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n (\langle w, x_i\rangle - \langle w_j, x_i\rangle)^2 &\le \max_{i\in[n]} \langle w - w_j, x_i\rangle^2 \\ &\le \max_{i\in[n]} \|w - w_j\|^2 \cdot \|x_i\|^2 \\ &\le \varepsilon^2.\end{aligned}$$

Hence,

$$\mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon) \le \mathcal{N}(\mathsf{B}_2^d, \|\cdot\|_2, \varepsilon). \tag{3.3}$$

In fact, a much stronger statement can be made: Since for any $x \in \mathcal{X}$

$$|\langle w, x\rangle - \langle w_j, x\rangle| \le \|w - w_j\| \, \|x\| \le \varepsilon,$$

the cover of the parameter space induces a cover of the function class *pointwise* (in the sup-norm $\|f - g\|_\infty = \sup_{x\in\mathcal{X}} |f(x) - g(x)|$) over the domain:

$$\mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon) \le \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \le \mathcal{N}(\mathsf{B}_2^d, \|\cdot\|_2, \varepsilon). \tag{3.4}$$

Recall that the covering number of B_2^d is

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

This gives a "parametric" growth of entropy

$$\log \mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon) \lesssim d \log(1 + 2/\varepsilon).$$

However, if d is large or infinite, this bound is loose. We will show that it also holds that

$$\log \mathcal{N}(\mathcal{F}, L^2(P_n), \varepsilon) \lesssim \varepsilon^{-2},$$

which is a nonparametric behavior. Hence, *the same class can be viewed as either parametric or nonparametric.* In fact, in the parametric behavior, it is not important that the domain of w is B_2^d since we would expect a similar estimate for other sets (including B_∞^d). In contrast, it will be crucial in nonparametric estimates that the norm of w is ℓ_2-bounded.

Jumping ahead, we will study neural networks and show a similar phenomenon: we can either count the number of neurons or connections (parameters) or we can calculate nonparametric "norm-based" estimates by looking at the norms of the layers in the network.

It's worth emphasizing again that (3.4) can lead to very loose bounds in high-dimensional situations. *A cover of function values on finite set of data can be significantly smaller than a cover with respect to sup norm.*

3.4 A more general result (Optional)

We have that for any fixed function

$$\mathbb{E}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n (f(X_i) - \mathbb{E}f(X))\right| \le \operatorname{var}(f)^{1/2} = \|f - \mathbb{E}f\|_{L^2(P)}.$$

Obviously this implies

$$\sup_{f\in\mathcal{F}} \mathbb{E}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n (f(X_i) - \mathbb{E}f(X))\right| \le \sup_{f\in\mathcal{F}} \operatorname{var}(f)^{1/2} =: \sigma$$

If we could ever prove

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n (f(X_i) - \mathbb{E}f(X))\right| \le C(\mathcal{F})\cdot\sigma,$$

it would imply that we only paid $C(\mathcal{F})$ for having a statement uniform in $f \in \mathcal{F}$.

Next, rather than assuming that functions in $\mathcal{F}$ are uniformly bounded, it will be enough to assume that they have an $L_2(P)$-integrable envelope F:

$$F(x) = \sup_{f\in\mathcal{F}} |f(x)|.$$

Rather than assuming that $F(x) \le 1$, we shall assume that $\|F\|_{L^2(P)}^2 = \mathbb{E}F(X)^2 \le \infty$ and everything will be phrased in terms of $\|F\|_{L^2(P)}^2$.

Now, let $H : [0,\infty) \mapsto [0,\infty)$ is such that $H(z)$ is non-decreasing for $z > 0$ and $z\sqrt{H(1/z)}$ is non-decreasing for $z \in (0,1]$. Assume

$$\int_0^D \sqrt{H(1/x)}dx \le C_H D\sqrt{H(1/D)}$$

for all $D \in (0,1]$, and suppose that

$$\sup_Q \log 2\mathcal{N}(\mathcal{F}, L^2(Q), \tau\|F\|_{L^2(Q)}) \le H(1/\tau)$$

for all $\tau > 0$. With this control on Koltchinskii-Pollard entropy, it follows that

$$\mathbb{E}\sup\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(X_i)-\mathbb{E}f(X))\right| \lesssim \sigma\sqrt{H\left(\frac{2\|F\|_{L^2(P)}}{\sigma}\right)} \tag{3.5}$$

if n is large enough. We refer to Giné & Nickl "Mathematical Foundations of Infinite-Dimensional Statistical Models" for more details, in particular Theorem 3.5.6 and the following corollaries.

Remarkably, under additional mild conditions on size of n, the inequality (3.5) can be reversed for a given P as soon as the entropy with respect to $L^2(P)$ indeed grows at least as $H\left(\frac{\|F\|_{L^2(P)}}{\sigma}\right)$.

Hence, the price we pay for uniformity in $f \in \mathcal{F}$ is truly

$$C(\mathcal{F}) \asymp \sqrt{H\left(\frac{\|F\|_{L^2(P)}}{\sigma}\right)}.$$

Of course, this expression is even simpler if $\sigma^2 = \sup_{f\in\mathcal{F}}\mathbb{E}(f(X)-\mathbb{E}f)^2$ is on the same order as $\|F\|^2_{L^2(P)} = \mathbb{E}\sup_f |f(X)|^2$.

4. COMBINATORIAL PARAMETERS

Let us gain some intuition for what can make $\widehat{\mathcal{R}}(\Theta)$ large. First, recall that

$$\widehat{\mathcal{R}}(\{\pm 1\}^n) = \mathbb{E}\sup_{\theta\in\{\pm1\}^n}\langle\theta,\epsilon\rangle = n.$$

Next, suppose that for $\alpha > 0$ and $v \in \mathbb{R}^n$,

$$\alpha\{\pm1\}^n + v \subseteq \Theta.$$

Then

$$\widehat{\mathcal{R}}(\Theta) \geq \widehat{\mathcal{R}}(\alpha\{\pm1\}^n + v) = \widehat{\mathcal{R}}(\alpha\{\pm1\}^n) = \alpha\widehat{\mathcal{R}}(\{\pm1\}^n) \geq \alpha n$$

Hence, "large cubes" inside Θ make Rademacher averages large. It turns out, this is the only reason $\widehat{\mathcal{R}}(\mathcal{F}|_{x_1,\ldots,x_n})$ can be large!

The key question is whether $\mathcal{F}|_{x_1,\ldots,x_n}$ contains large cubes for a given class $\mathcal{F}$.

4.1 Binary-Valued Functions

Let's start with function classes of $\{0,1\}$-valued functions. In this case, $\mathcal{F}_{x_1,\ldots,x_n}$ is either a full $\{0,1\}^n$ cube or not. Consider the particular example of threshold functions on the real line. Take any point x_1. Clearly, $\mathcal{F}|_{x_1} = \{0,1\}$, which is a one-dimensional cube. Take two points x_1, x_2. We can only realize sign patters $(0,0),(0,1),(1,1)$, but not $(1,0)$. Hence, for no two points can we get a cube.

Definition: Let $\mathcal{F} = \{f : \mathcal{X} \to \{0,1\}\}$. We say that $\mathcal{F}$ shatters $x_1,\ldots,x_n \in \mathcal{X}$ if

$\mathcal{F}|_{x_1,\ldots,x_n} = \{0,1\}^n$. The Vapnik-Chervonenkis dimension of $\mathcal{F}$ is

$$\mathrm{vc}(\mathcal{F}) = \max\{n : \mathcal{F} \text{ shatters some } \ x_1, \ldots, x_n\}$$

Lemma (Sauer-Shelah-Vapnik-Chervonenkis): If $\mathrm{vc}(\mathcal{F}) = d < \infty$,

$$\mathrm{card}\left(\mathcal{F}|_{x_1,\ldots,x_n}\right) \leq \sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^d$$

This result is quite remarkable. It says that as soon as $n > \mathrm{vc}(\mathcal{F})$, the proportion of the cube that can be realized by $\mathcal{F}$ becomes very small (n^d vs 2^n). This combinatorial result is at the heart of empirical process theory and the early developments in pattern recognition.

In particular, the lemma can be interpreted as a covering number upper bound:

$$\mathcal{N}(\mathcal{F}, L^\infty(P_n), \varepsilon) \leq \left(\frac{en}{d}\right)^d$$

for any $\epsilon > 0$. Observe that these numbers are with respect to $L^\infty(P_n)$ rather than $L^2(P_n)$, and hence can be an overkill. Indeed, $L^\infty(P_n)$ covering numbers are necessarily n-dependent while we can hope to get dimension-independent $L^2(P_n)$ covering numbers. Indeed, this result (Dudley, Haussler) was already mentioned: for a binary-valued class with finite $\mathrm{vc}(\mathcal{F}) = d$,

$$\mathcal{N}(\mathcal{F}, L^2(P_n), \epsilon) \lesssim \left(\frac{C}{\epsilon}\right)^{Cd}.$$

Hence, a class with finite VC dimension is "parametric". On the other hand, if $\mathrm{vc}(\mathcal{F})$ is infinite, then $\mathcal{F}|_{x_1,\ldots,x_n}$ is a full cube for arbitrarily large n (for some appropriately chosen points). Hence, Rademacher averages of this set are too large and there is no uniform convergence for all P (to see this, consider P supported on the shattered set). Hence, finiteness of VC dimension is a characterization (of both distribution-free learnability and uniform convergence).

4.2 Real-Valued Functions

For binary-valued functions, the size of the cube contained in $\mathcal{F}|_{x_1,\ldots,x_n}$ was trivially 1, and we only varied n to see where the phase transition occurs. In contrast, for a general real-valued function class, it is feasible that $\mathcal{F}|_{x_1,\ldots,x_n}$ contains a cube of size α, but not larger than α; this extra parameter is in addition to the dimensionality of the cube. To deal with this extra degree of freedom, we fix the scale α and ask for the largest size n such that $\mathcal{F}|_{x_1,\ldots,x_n}$ contains a (translate of a) cube of size α. A true containment statement would read $s + (\alpha/2)\{-1,1\}^n \subseteq \mathcal{F}|_{x_1,\ldots,x_n}$. However, it is enough to ask that the equalities for the vertices are replaced with inequalities:

Definition: We say that $\mathcal{F}$ *shatters* a set of points $x_1, \ldots, x_n$ at scale α if there exists

$s \in \mathbb{R}^n$ such that

$$\forall \epsilon \in \{\pm 1\}^n, \exists f \in \mathcal{F} \text{ s.t. } \begin{cases} f(x_t) \geq s_t + \alpha/2 & \text{if } \epsilon = +1 \\ f(x_t) \leq s_t - \alpha/2 & \text{if } \epsilon = -1 \end{cases}$$

The combinatorial dimension $\mathrm{vc}(\mathcal{F}, \alpha)$ of $\mathcal{F}$ (on domain $\mathcal{X}$) at scale α is defined as the size n of the largest shattered set.

4.2.1 Example: non-decreasing functions

Consider the class of nondecreasing functions $f : \mathbb{R} \to [0, 1]$. First, observe that a pointwise cover of this class does not exist ($\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \epsilon) = \infty$ for any $\epsilon < 1/2$). However, $\mathcal{N}(\mathcal{F}, L^\infty(P_n), \epsilon)$ is necessarily finite. Let's calculate the scale-sensitive dimension of this class.

Claim: $\mathrm{vc}(\mathcal{F}, \epsilon) \leq \epsilon^{-1}$. Indeed, fix any $x_1, \ldots, x_n$ and assume these are arranged in an increasing order. Suppose $\mathcal{F}$ shatters this set. Take the alternating sequence $\epsilon = (+1, -1, \ldots)$. We then must have a nondecreasing function that is at least $s_1 + \alpha/2$ at x_1 but then no greater than $s_2 - \alpha/2$ at x_2. The nondecreasing constraint implies that $s_2 \geq s_1 + \alpha$. A similar argument then holds for the next point and so forth. Since functions are bounded, $n\alpha \leq 1$, which concludes the proof.

4.2.2 Control of covering numbers

The following generalization of the earlier result for binary-valued functions is due to Mendelson and Vershynin:

Theorem: Let $\mathcal{F}$ be a class of functions $\mathcal{X} \to [-1, 1]$. Then for any distribution P,

$$\mathcal{N}(\mathcal{F}, L_2(P), \varepsilon) \leq \left(\frac{c}{\varepsilon}\right)^{c \cdot \mathrm{vc}(\mathcal{F}, \varepsilon/c)}$$

for all $\epsilon > 0$. Here c is an absolute constant.

In particular, plugging into the entropy integral yields

$$\int \sqrt{\mathrm{vc}(\mathcal{F}, \varepsilon) \log(1/\varepsilon)} d\varepsilon$$

Rudelson-Vershynin: $\log(1/\epsilon)$ can be removed.

Back to the class of non-decreasing functions, we immediately get

$$\log \mathcal{N}(\mathcal{F}, L_2(P_n), \varepsilon) \lesssim \varepsilon^{-1} \cdot \log\left(\frac{c}{\varepsilon}\right).$$

In particular, Rademacher averages of this class scale as $n^{-1/2}$ since this is a nonparametric class with entropy exponent $p < 2$.

4.3 Scale-sensitive dimension of linear class via Perceptron

In this section, we will prove that

Proposition: For

$$\mathcal{F} = \{x \mapsto \langle w, x\rangle : w \in \mathsf{B}_2^d\}$$

and $\mathcal{X} \subseteq \mathsf{B}_2^d$, it holds that

$$\mathrm{vc}(\mathcal{F}, \alpha) \lesssim 16\alpha^{-2}.$$

We turn to the Perceptron algorithm, defined as follows. We start with $\widehat{w}_0 = 0$. At time $t = 1, \ldots, T$, we observe $x_t \in \mathcal{X}$ and predict $\widehat{y}_t = \mathrm{sign}(\langle \widehat{w}_t, x_t\rangle)$, a *deterministic* guess of the label of x_t given the hypothesis $\widehat{w}_t$. We then observe the true label of the example $y_t \in \{\pm 1\}$. If $\widehat{y}_t \neq y_t$, we update

$$\widehat{w}_{t+1} = \widehat{w}_t + y_t x_t,$$

and otherwise $\widehat{w}_{t+1} = \widehat{w}_t$.

Lemma (Novikoff'62): For any sequence $(x_1, y_1), \ldots, (x_T, y_T) \in \mathsf{B}_2^d \times \{\pm 1\}$ the Perceptron algorithm makes at most γ^{-2} mistakes, where γ is the margin of the sequence, defined as

$$\gamma = \max_{w^* \in \mathsf{B}_2^d} \min_t y_t \langle w^*, x_t\rangle$$

Proof. If a mistake is made on round t,

$$\|\widehat{w}_{t+1}\|^2 = \|\widehat{w}_t + y_t x_t\|^2 \leq \|\widehat{w}_t\|^2 + 2y_t\langle \widehat{w}_t, x_t\rangle + 1 \leq \|\widehat{w}_t\|^2 + 1$$

Denote the number of mistakes at the end as m. Then $\|\widehat{w}_T\|^2 \leq m$. Next, for w^*,

$$\gamma \leq \langle w^*, y_t x_t\rangle = \langle w^*, \widehat{w}_{t+1} - \widehat{w}_t\rangle,$$

and so by summing and telescoping, $m\gamma \leq \langle w^*, \widehat{w}_T\rangle \leq \sqrt{m}$. This concludes the proof. □

Remarkably, the number of mistakes does not depend on the dimension d. We will now show that the mistake bound translates into a bound on the scale-sensitive dimension.

Proof of Proposition. Suppose there exist a shattered set $x_1, \ldots, x_m \in \mathsf{B}_2^d$: there exists $s_1, \ldots, s_m \in [-1, 1]$ such that for any sequence of signs $\epsilon = (\epsilon_1, \ldots, \epsilon_m)$ there exists a $w_\epsilon \in \mathsf{B}_2^d$ such that

$$\epsilon_i(\langle w_\epsilon, x_i\rangle - s_i) \geq \alpha/2.$$

Claim: we can reparametrize the problem so that $s_i = 0$. Indeed, take

$$\tilde{w}_\epsilon = [w_\epsilon, 1], \quad \tilde{x}_i = [x_i, -s_i].$$

Then we have

$$\epsilon_i\langle \tilde{w}_\epsilon, \tilde{x}_i\rangle \geq \alpha/2.$$

while the norms are at most $\sqrt{2}$:

$$\|\tilde{w}_\epsilon\|^2 = \|w_\epsilon\|^2 + 1 \leq 2, \qquad \|\tilde{x}_i\|^2 \leq 2$$

Now comes the key step. We run Perceptron on the sequence $\tilde{x}_1/\sqrt{2}, \ldots, \tilde{x}_m/\sqrt{2}$ and $y_i = -\widehat{y}_i$. That is, we force Perceptron to make mistakes on every round, no matter what the predictions are. It is important that Perceptron makes deterministic predictions for this argument to work. Note that the sequence of predictions of Perceptron defines the sequence $y = (y_1, \ldots, y_n)$ with

$$y_i \langle \tilde{w}_y/\sqrt{2}, \tilde{x}_i/\sqrt{2} \rangle \geq \alpha/4.$$

Hence, by Novikoff's result,

$$m \leq 16/\alpha^2.$$

□

Interestingly, both Perceptron and VC theory were developed in the 60's as distinct approaches (online vs batch), yet the connection between them runs deeper than was recognized, until recently. In particular, the above proof in fact shows that a stronger *sequential* version of $\mathrm{vc}(\mathcal{F}, \alpha)$ is also bounded by $16\alpha^{-2}$, where (roughly speaking) sequential analogues allow the sequence to evolve as a predictable process with respect to a dyadic filtration. It turns out that there are sequential analogues of Rademacher averages, covering numbers, Dudley chaining, and combinatorial dimensions, and these govern *online* (rather than i.i.d.) learning. If there is time, we will mention these towards the end of the course.