MATH 281C: Mathematical Statistics

Lecture 6

In the previous class, we presented an important lemma, which allows us to show that VC classes have polynomial discrimination and then to control their Rademacher complexity accordingly. The first goal of this lecture is to provide a proof of this lemma.

1 Proof of the Sauer-Shelah-Vapnik-Chervonenkis Lemma

Lemma 1.1 (Sauer-Shelah-Vapnik-Chervonenkis)**.** Suppose that the the VC dimension of a Boolean class $\mathcal{F}$ of functions on $\mathcal{X}$ is D. Then for every $n \geq 1$ and $x_1, \ldots, x_n \in \mathcal{X}$, we have

$$|\mathcal{F}(x_1, \ldots, x_n)| \leq \sum_{\ell=0}^{D} \binom{n}{\ell}.$$

Here $\binom{n}{k}$ is taken to be 0 if $n < k$. Moreover, if $n \geq D$, then

$$|\mathcal{F}(x_1, \ldots, x_n)| \leq \sum_{\ell=0}^{D} \binom{n}{\ell} \leq \left(\frac{en}{D}\right)^D.$$

Proof. Let us briefly review the notation and setup. We have a Boolean class $\mathcal{F}$ with VC dimension D, and we consider some fixed $n \geq 1$. For notational simplicity, we let $\Delta = \mathcal{F}(x_1, \ldots, x_n)$. Notice that Δ is a set, but it can equally be represented as a boolean matrix. Remember that elements of Δ are bit vectors in $\{0,1\}^n$, so if we write those vectors out as u_j for $j = 1, \ldots, M := |\Delta|$, we can represent Δ as an $n \times M$ matrix whose j-th column is u_j. By construction, Δ contains no duplicate columns, since u_j's represent the distinct elements of the set Δ.

As usual with VC stuff, we will want to reason about subsets of $\{x_1, \ldots, x_n\}$. We will interchangeably think of such subsets as subsets of $\{1, \ldots, n\}$ when convenient. For any such subset S, we let Δ_S denote the $|S| \times M$ submatrix of Δ that keeps only the rows corresponding to S. For example, if $S = \{x_1, x_5, x_8\}$, then Δ_S is the $3 \times M$ submatrix of Δ consisting of only the first, fifth and eighth rows of Δ.

It is important to notice, the statement that S is shattered by $\mathcal{F}$ is equivalent to the statement that every element of $\{0,1\}^{|S|}$ appears as a column of Δ_S. Using that second definition, we can say that given any boolean matrix $\widetilde{\Delta}$, $\widetilde{\Delta}$ shatters a subset S of the indices if every element of $\{0,1\}^{|S|}$ appears as a column of $\widetilde{\Delta}_S$.

Now, since the VC dimension of $\mathcal{F}$ is D, no set S with $|S| > D$ can be shattered. Therefore, the number of subsets of $\{x_1, \ldots, x_n\}$ that can be shattered by $\mathcal{F}$ is bounded above by

$$\sum_{\ell=0}^{D} \binom{n}{\ell}.$$

We therefore have to show that the number of columns of Δ is at most the number of subsets of $\{x_1, \ldots, x_n\}$ that can be shattered by Δ. Denote the latter by $\mathcal{S}(\Delta)$.

We claim that

$$M \leq \mathcal{S}(\Delta).$$

The proof of this follows an idea called *downshifting*. We can make this more precise by defining the transformed matrix Δ' as follows, assuming that we are downshifting using the first row $i_0 = 1$:

$$\Delta'_{ij} = \begin{cases} \Delta_{ij} & \text{if } i \neq i_0, \\ \Delta_{i_0 j} & \text{if } \Delta_{i_0 j} = 0 \text{ or } \Delta_{\cdot j} = (1, u) \text{ and } (0, u) \text{ apprears as a column of } \Delta, \\ 0 & \text{otherwise.} \end{cases}$$

That is, in the i_0-th row, let 0's remain still, and replace all 1's with 0's unless this produces a duplicate column in Δ. We then claim that

$$\mathcal{S}(\Delta') \leq \mathcal{S}(\Delta). \tag{6.1}$$

This claim is the key component of the proof.

To prove (6.1), we need to show that any subset S that Δ does not shatter is also not shattered by Δ'. Fix an arbitrary S that is not shattered by Δ, and choose the i_0-th row ($i_0 \in S$) for the downshift. Without loss of generality, assume $i_0 = 1$. Then there exists an element of $u = (u_1, \ldots, u_{|S|}) \in \{0, 1\}^{|S|}$ that does not appear as a column in Δ_S.

- If $u = (1, v)$, then u does not appear as a column of Δ'_S either; otherwise u mush be in Δ_S because downshifting never replace a 0 by a 1.
- If $u = (0, v)$, consider two cases: (i) if $u' = (1, v)$ did not appear in Δ_S either, then u is not a column of Δ'_S; otherwise by downshifting, one of u and u' has to be a column of Δ'_S. (ii) if $u' = (1, v)$ does appear in Δ_S, then since $(0, v)$ is not, downshifting replaces $(1, v)$ by $(0, v)$. Hence $(1, v) \in \{0, 1\}^{|S|}$ cannot be a column of Δ'_S, so that Δ' does not shatter $\mathcal{S}$.

Combining the pieces completes the proof of (6.1).

Repeat downshifting until we cannot downshift anymore, and let the Δ^* be the resulting matrix. By (6.1), $\mathcal{S}(\Delta^*) \leq \mathcal{S}(\Delta)$. Our final step is therefore to show $M \leq \mathcal{S}(\Delta^*)$.

The proof of this claim proceeds by constructing, for each column j, a different set S_j (i.e., a subset of $\{x_1, \ldots, x_n\}$) that Δ^* shatters. For $j = 1, \ldots, M$,

$$S_j = \{1 \leq i \leq n : \Delta^*_{ij} = 1\}. \tag{6.2}$$

These sets are distinct because if $S_j = S_{j'}$, then column j and column j' have 1's at the same location and therefore are duplicates of each other, which is ruled out because downshifting preserves the distinctness of columns. So we only need to show that S_j is shattered by Δ^*. For that, we first notice that $\Delta^*_{S_j}$ has the all 1's vector (the vector that has 1 at each coordinate) as its j-th column. Next, suppose there was some vector $u \in \{0, 1\}^{|S_j|}$ having exactly one zero that was NOT a column of $\Delta^*_{S_j}$. Then we could downshift using the row where this zero appears, so that the j-th column would have turned into u, contradicting the minimality of Δ^*. We hence conclude that every pattern

of 0's and 1's which has only one zero must appear in $\Delta^*_{S_j}$. An inductive argument then shows that every pattern of 0's and 1's (i.e. every $u \in \{0,1\}^{|S_j|}$) must appear in $\Delta^*_{S_j}$. Therefore, Δ^* shatters S_j for $j = 1, \ldots, M$, implying $M \le \mathcal{S}(\Delta^*)$.

Finally, it remains to show that when $n \ge D$,

$$\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{D} \le \left(\frac{en}{D}\right)^D.$$

To see this, let B be a binomial random variable that corresponds to n tosses with probability of success $1/2$, that is, $B = \sum_{i=1}^n b_i$ with b_i iid having Bernoulli$(1/2)$. Then the LHS above equals

$$2^n \mathbb{P}(B \le D) = 2^n \sum_{\ell=0}^{D} \mathbb{P}(B = \ell).$$

On the other hand, note that $I(B \le D) \le (D/n)^{B-D}$ (recall that $D/n \le 1$), from which it follows that

$$\begin{aligned} 2^n \mathbb{P}(B \le D) \le 2^n \mathbb{E}(D/n)^{B-D} &= (D/n)^{-D} 2^n \mathbb{E}(D/n)^B \\ &= \left(\frac{n}{D}\right)^D \sum_{\ell=0}^{n} \left(\frac{D}{n}\right)^\ell \binom{n}{\ell} = \left(\frac{n}{D}\right)^D (1 + D/n)^n \le \left(\frac{en}{D}\right)^D. \end{aligned}$$

The proof of Lemma 1.1 is now complete. □

2 Covering and Packing Numbers

As mentioned in the previous lecture, chaining gives sharper bounds for $R_n(\mathcal{F}(x_1, \ldots, x_n))$ compared to the simple bounds discussed in the previous lecture. In order to discuss chaining, we need to be familiar with the notions of *covering* and *packing numbers*.

Let T be a set equipped with a pseudometric d. A pseudometric is a map $d : T \times T \to [0, \infty)$ that satisfies

- $d(t, t) = 0$;
- $d(t_1, t_2) = d(t_2, t_1)$;
- $d(t_1, t_2) \le d(t_1, t_3) + d(t_3, t_2)$ for all $t_1, t_2, t_3 \in T$.

If, in addition, it also satisfies $d(t_1, t_2) > 0$ for $t_1 \ne t_2 \in T$, then it is called a metric.

Example 2.1. Given a space of functions $\mathcal{F}$ mapping $\mathcal{X} \to \mathbb{R}$, we can define a pseudometric by

$$(f, g) \mapsto \sqrt{\frac{1}{n} \sum_{i=1}^{n} \{f(x_i) - g(x_i)\}^2}.$$

It will not, however, be a metric in general for arbitrary fixed $x_1, \ldots, x_n \in \mathcal{X}$. We will see more concrete examples shortly.

Definition 2.1 (δ-net)**.** Let $F \subseteq T$ and $\delta > 0$. We say a subset $\{t_1, \ldots, t_N\} \subseteq T$ is a δ-net of F if for every $t \in F$, there exists a $t_i \in F$ such that $d(t, t_i) \le \delta$.

Definition 2.2 (Covering number)**.** The δ-covering number $F \subseteq T$ is denoted by $N_T(\delta, F, d)$ and is defined as the size of the minimal δ-net of F. That is,

$$N_T(\delta, F, d) = \min\{|\mathcal{N}_\delta| : \mathcal{N}_\delta \text{ is a } \delta\text{-net of } F\}.$$

The logarithm of this quantity is called the *metric entropy*. If $N_T(\delta, T, d) < \infty$ for every $\delta > 0$, we say that T is *totally bounded*.

Remark 2.1. The centers need not lie inside the subset F, hence the need to include the T subscript. We can regard F as a (pseudo)metric space in its own right, and define $N_F(\delta, F, d)$ accordingly. These different concepts of covering number are closely related. In fact,

$$N_F(2\delta, F, d) \leq N_T(\delta, F, d) \leq N_F(\delta, F, d).$$

We leave the proof of this property as an exercise.

The notion of covering numbers is closely related to that of packing numbers which are defined next.

Definition 2.3 (δ-packing)**.** For $\delta > 0$, we say that a subset $\{t_1, \ldots, t_M\} \subseteq F$ is a *δ-packing* if for every $i \neq j$, $d(t_i, t_j) > \delta$.

Definition 2.4 (Packing number)**.** The *δ-packing number* $M(\delta, F, d)$ of F is defined as the largest M such that there exists a δ-packing of F with M elements. That is,

$$M(\delta, F, d) = \max\{|\mathcal{P}_\delta| : \mathcal{P}_\delta \text{ is a } \delta\text{-packing of } F\}.$$

The following result shows that covering and packing numbers are closely related to each other.

Lemma 2.1. For every $\delta > 0$, we have

$$N_F(\delta, F, d) \leq M(\delta, F, d) \leq N_T(\delta/2, F, d) \leq N_F(\delta/2, F, d). \tag{6.3}$$

A basic case we need to understand in order to make use of packings and coverings is the case of $\mathbb{R}^k$. This setting is in some sense a limiting simple case for the concepts we have considered here. More generally, we will be interested in analyzing covering and packing numbers of function classes. When these classes are parametric, however, they will turn out to be have roughly like k-dimensional Euclidean space in that $N(\epsilon, F, d) \approx \epsilon^{-k}$. On the other hand, some classes will turn out to be nonparametric and will exhibit scalings like $\exp(\epsilon^{-k})$.

2.1 Parametric classes

Proposition 2.1. Suppose $\|\cdot\|$ denotes any norm in $\mathbb{R}^k$. For example, it might be the usual Euclidean norm or the ℓ_1-norm, $\|x\|_1 = \sum_{j=1}^k |x_j|$. Let

$$B_R = \{x \in \mathbb{R}^k : \|x\| \leq R\}.$$

Then, for every $\epsilon > 0$, we have

$$M(\epsilon R, B_R, d) \leq \left(1 + \frac{2}{\epsilon}\right)^k, \tag{6.4}$$

where d denotes the metric corresponding to the norm $\|\cdot\|$.

Proof. Let $x_1, \ldots, x_N$ be any set of points in B_R that is ϵR-separated, i.e., $\|x_i - x_j\| > \epsilon R$ for all $i \neq j$. Then the following closed balls

$$B(x_i, \epsilon R/2) := \{x \in \mathbb{R}^k : \|x - x_i\| \leq \epsilon R/2\}, \quad i = 1, \ldots, N,$$

are disjoint. Moreover, $B(x_i, \epsilon R/2) \subseteq B_{R+\epsilon R/2}$. As a result,

$$\sum_{i=1}^{N} \mathrm{Vol}(B(x_i, \epsilon R/2)) \leq \mathrm{Vol}(B_{R+\epsilon R/2}).$$

If we let Λ denote the volume of the unit ball B_1, then the above inequality becomes

$$\sum_{i=1}^{N} \left(\frac{\epsilon R}{2}\right)^k \Lambda \leq \left(R + \frac{\epsilon R}{2}\right)^k \Lambda,$$

which immediately proves (6.4). □

The argument used above to prove (6.4) is known as the **volumetric argument** because it is based on a volume comparison.

Example 2.2 (A simple parametric case)**.** Let $\mathcal{F}$ denote the set of functions $\{x \mapsto \langle \beta, x \rangle : \beta \in \mathbb{R}^k\}$, and we equip it with a pseudometric

$$d(\beta, \gamma) = \sqrt{\mathbb{E}_Q \langle \beta - \gamma, X \rangle^2}$$

for some distribution Q on $\mathbb{R}^k$. Since

$$d(\beta, \gamma) \leq \underbrace{\sqrt{\mathbb{E}_Q \|X\|_2^2}}_{=:\|X\|_Q} \cdot \|\beta - \gamma\|_2,$$

we should expect $N(\delta, \mathcal{F}, d) \approx \delta^{-k}$ and in fact this is true.

The following proposition gives a general result for a particular type of parametric class of functions, extending the example presented above. Note that when D and $\|\Gamma\|_Q$ below are constants, the covering number bound given by the result below is of the form ϵ^{-k}.

Proposition 2.2. Let $\Theta \subseteq \mathbb{R}^k$ be a non-empty bounded subset with Euclidean diameter D, and let $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ be a class of functions on $\mathcal{X}$ indexed by Θ such that for some nonnegative function $\Gamma : \mathcal{X} \to \mathbb{R}$,

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq \Gamma(x) \cdot \|\theta_1 - \theta_2\| \tag{6.5}$$

for all $x \in \mathcal{X}$ and $\theta_1, \theta_2 \in \Theta$. Here $\|\cdot\|$ denotes the usual Euclidean norm on $\mathbb{R}^k$.

Fix a probability measure Q on $\mathcal{X}$, and let d denote the pseudometric on $\mathcal{F}$ defined by

$$d(f, g) = \sqrt{\int_{\mathcal{X}} \{f(x) - g(x)\}^2 \mathrm{d}Q(x)} = \sqrt{\mathbb{E}_Q \{f(X) - g(X)\}^2}.$$

Then, for every $\epsilon > 0$,

$$M(\epsilon, \mathcal{F}, d) \leq \left(1 + \frac{2D\|\Gamma\|_Q}{\epsilon}\right)^k \quad \text{with} \quad \|\Gamma\|_Q = \left\{\int_{\mathcal{X}} |\Gamma(x)|^2 \mathrm{d}Q(x)\right\}^{1/2}.$$

Proof. Condition (6.5) implies that for every $\theta_1, \theta_2 \in \Theta$,

$$d(f_{\theta_1}, f_{\theta_2}) \leq \|\theta_1 - \theta_2\| \cdot \|\Gamma\|_Q.$$

As a result, every ϵ-separated subset of $\mathcal{F}$ under metric d is automatically an $\epsilon/\|\Gamma\|_Q$-separated subset on Θ. Consequently,

$$M(\epsilon, \mathcal{F}, d) \leq M\left(\frac{\epsilon}{\|\Gamma\|_Q}, \Theta, \|\cdot\|\right).$$

To bound the Euclidean packing number, we use the assumption that Θ has diameter $\leq D$ so that $\Theta \subseteq B(a, D)$ for every $a \in \Theta$. Pick a arbitrary $a \in \Theta$, we have

$$M\left(\frac{\epsilon}{\|\Gamma\|_Q}, \Theta, \|\cdot\|\right) \leq M\left(\frac{\epsilon}{\|\Gamma\|_Q}, B(a, D), \|\cdot\|\right).$$

To bound the RHS, using Proposition 2.1 (note that we can take $a = 0$ above because balls of the same radius will have the same packing numbers regardless of their center) yields

$$M\left(\frac{\epsilon}{\|\Gamma\|_Q}, B(a, D), \|\cdot\|\right) \leq \left(1 + \frac{2D}{\epsilon}\|\Gamma\|_Q\right)^k,$$

which completes the proof. □

2.2 Nonparametric classes

The difficulty of estimating a nonparametric function depends on the structure or landscape of this function. Quantitatively, this is reflected by the convergence rate, which is often determined by the global metric entropy over the whole function class (or over large subsets of it). The advantage is that the metric entropies are available in approximation theory for many function classes; see Lorentz, Golitschek and Markovoz (1996) and Section 2 of van der Vaart and Wellner (1996). The most standard examples of nonparametric function classes are smoothness classes. These will have covering numbers that are exponential in $1/\epsilon$. We will introduce smoothness and convex classes and describe their covering numbers in one dimension, and then generalize to multiple dimensions.

2.2.1 Smoothness classes

Throughout this section, we define the pseudometric as

$$d(f, g) = \sup_{x \in [0,1]} |f(x) - g(x)|.$$

Fix $\alpha > 0$, let β be the largest integer that is strictly smaller than α. Then, define the function class $\mathcal{S}_\alpha$ that consists of functions f on $[0, 1]$ satisfying the following properties:

(i) f is continuous on $[0, 1]$;

(ii) f is β times differentiable on $(0, 1)$;

(iii) $|f^{(k)}(x)| \leq 1$ (or by some constant $C_0 > 0$) for all $k = 0, 1, \ldots, \beta$ and $x \in [0, 1]$, where $f^{(0)}(x) = f(x)$;

(iv) $|f^{(\beta)}(x) - f^{(\beta)}(y)| \leq |x-y|^{\alpha-\beta}$ (or $\leq L|x-y|^{\alpha-\beta}$) for all $x, y \in (0,1)$.

Note that if $\alpha = 1$, $\mathcal{S}_\alpha$ is the class of differentiable 1-Lipschitz functions on $[0,1]$ that are bounded by 1. In general, α measures the degree of smoothness, with the gap between β and α measuring specifically the smoothness of the β-th derivative.

Theorem 2.1. There exist positive constants C_1 and C_2 depending on α alone such that for all $\epsilon \in (0,1)$,

$$C_1 \epsilon^{-1/\alpha} \leq \log M(\epsilon, \mathcal{S}_\alpha, \rho) \leq C_2 \epsilon^{-1/\alpha}.$$

Thus the ϵ-metric entropy of the smoothness class $\mathcal{S}_\alpha$ in one dimension grows as $\epsilon^{-1/\alpha}$. Here α denotes the degree of smoothness (the higher α is, the smoother the functions in $\mathcal{S}_\alpha$).

This result has a direct generalization to multiple dimensions. As before, $\alpha > 0$, and β is the largest integer that is strictly smaller than α. For a vector $p = (p_1, \ldots, p_d)$ consisting of nonnegative integers $p_1, \ldots, p_d$, let $\langle p \rangle = p_1 + \cdots + p_d$. Define the partial derivative operator

$$D^p = \partial^{\langle p \rangle} / \partial x_1^{p_1} \cdots \partial x_d^{p_d}.$$

The class $\mathcal{S}_{\alpha,d}$ is defined to consist of all functions f on $[0,1]^d$ that satisfy

(i) f is continuous on $[0,1]^d$;

(ii) All partial derivatives D^p of f exist on $(0,1)^d$ for $\langle p \rangle \leq \beta$;

(iii) $|D^p f(x)| \leq 1$ for all p with $\langle p \rangle \leq \beta$ and $x \in [0,1]^d$;

(iv) $|D^p f(x) - D^p f(y)| \leq |x-y|^{\alpha-\beta}$ for all p with $\langle p \rangle = \beta$ and $x, y \in (0,1)^d$.

Once again, we consider the supremum metric. Note that in this case, setting $\alpha = 1$ gives the set of differentiable 1-Lipschitz functions with respect to the L_2 metric that are also bounded by 1. Below are some bounds on metric entropies proved by Kolmogorov. See Section 2.7 of van der Vaart and Wellner (1996).

Theorem 2.2. There exist positive constants C_1 and C_2 depending only on (d, α) such that for all sufficiently small $\epsilon > 0$,

$$C_1 \epsilon^{-d/\alpha} \leq \log M(\epsilon, \mathcal{S}_{\alpha,d}, \rho) \leq C_2 \epsilon^{-d/\alpha}.$$

Thus the metric entropy of a smoothness class of functions with smoothness α and dimension d scales as $\epsilon^{-d/\alpha}$. This grows as d increases and goes down as α increases.

2.2.2 Monotone classes

Let $\mathcal{M}$ denote the class of all functions f on $[0,1]$ such that

(i) f is non-decreasing on $[0,1]$;

(ii) $|f(x)| \leq 1$ for all $x \in [0,1]$.

For a probability measure Q on $[0,1]$, let ρ_Q denote the metric on $\mathcal{M}$ given by

$$\rho_Q(f,g) = \|f-g\|_Q = \left\{\int \{f(x)-g(x)\}^2 \mathrm{d}Q(x)\right\}^{1/2}.$$

Then it can be proved that (see Section 2.7.2 in van der Vaart and Wellner (1996))

$$\log M(\epsilon, \mathcal{M}, \rho_Q) \leq \frac{C}{\epsilon}$$

for every probability measure Q on $[0,1]$, where $C > 0$ is an absolute constant. There exist probability measures Q for which a lower bound of C_2/ϵ also holds on the metric entropy. Comparing this result with Theorem 2.1, it is clear that the covering numbers of $\mathcal{M}$ (class of bounded monotone functions) are comparable to the smoothness class $\mathcal{S}_1$ (class of bounded Lipschitz continuous functions). Thus bounded monotone functions have the same metric entropy as bounded Lipschitz functions even though monotone functions need not be continuous.

2.2.3 Lipschitz convex classes

Let $C(L)$ denote the class of all functions f on $[0,1]^d$ such that

(i) f is convex on $[0,1]^d$;

(ii) $|f(x)| \leq 1$ for all $x \in [0,1]^d$;

(iii) $|f(x) - f(y)| \leq L\|x-y\|$.

Let ρ be the supremum metric on $[0,1]^d$. Bronshtein (1976) proved that for ϵ sufficiently small,

$$C_1\left(\frac{L}{\epsilon}\right)^{d/2} \leq \log N(\epsilon, C(L), \rho) \leq C_2\left(\frac{1+L}{\epsilon}\right)^{d/2},$$

where $C_1, C_2 > 0$ are constants not depending on ϵ. Comparing this to Theorem 2.2, it is clear that, in terms of metric entropy, $C := C(1)$ is comparable to the smoothness class $\mathcal{S}_{2,d}$. This is interesting because convex functions are not necessarily twice differentiable in the usual sense. Yet, they possess the regularity of second order smoothness in terms of metric entropy.

For more results on covering numbers for convex functions, see Guntuboyina and Sen (2013).

References

Bronshtein, E. M. (1976). ϵ-entropy of convex sets and functions. *Siberian Math. J.* **17** 393–398.

Guntuboyina, A. and Sen, B. (2013). Covering number for convex functions. *IEEE Transactions on Information Theory* **59** 1957–1965.

Lorentz, G. G., Golitschek, M. V. and Markovoz, Y. (1996). *Constructive Approximation: Advanced Problems.* Springer, New York.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer-Verlag, New York.