Lecture 6: Covering Numbers and Chaining

Akshay Krishnamurthy
akshay@cs.umass.edu

September 21, 2017

1 Recap

We have been discussing different ways to prove uniform convergence, which we earlier saw was important for obtaining excess risk or sample complexity bounds. Today we will see our last and most powerful technique. Recall the definitions of VC-dimension and Rademacher complexity. $\text{VCdim}(\mathcal{H})$ is the size of the largest point set that can be shattered by $\mathcal{H}$ and $\mathcal{R}(A) = \mathbb{E}_\sigma \sup_{a \in A} \frac{1}{n}\langle \sigma, a \rangle$ for vectors $A \subset \mathbb{R}^n$. The two results we saw before were

$$\mathbb{E}_{Z^n} \sup_{f \in \mathcal{F}} R(f) - \hat{R}(f) \leq \begin{cases} 2\mathbb{E}_{Z^n}\mathcal{R}(\ell \circ \mathcal{F} \circ Z^n) & \text{Rademacher bound} \\ \sqrt{\frac{2d\log(2en/d))}{n}} & \text{VC bound} \end{cases}$$

Both bounds here are statements about uniform convergence, which we saw can be directly translated to sample complexity bounds by application of McDiarmid's inequality.

Today we will study another way to analyze the Rademacher complexity purely in geometric terms, which can be much easier to think about conceptually and work with analytically. This leads to covering number bounds and a powerful uniform convergence statement, called Dudley's entropy integral.

2 Covering Number bounds

Let $A \subset \mathbb{R}^n$ be a set of vectors. We want to obtain bounds on $\mathcal{R}(A)$. If A is finite, then we already know what to do, since we can "apply union bound," which we did last time in the Massart Lemma

Lemma 1 (Massart Lemma)**.** *If $A = \{a_1, \ldots, a_N\}$ is a finite set with $\bar{a} = \frac{1}{N}\sum_i a_i$, then $\mathcal{R}(A) \leq \max_a \|a - \bar{a}\|\sqrt{2\log N}/n$.*

If A is not finite, then a natural idea is to try to discretize the set. This is precisely the idea with covering numbers. The basic idea is to find a finite subset $\tilde{A} \subset A$ such that for every $a \in A$, there is some $\tilde{a} \in \tilde{A}$ that is quite close to a. Then we apply Massart Lemma on $\tilde{A}$ and pay a little bit extra depending on $\|a - \tilde{a}\|$.

To state things precisely we need some definitions.

Definition 2 (Metric space)**.** *A metric space (S, ρ) consists of a set of objects S, and a function $\rho : S \times S \Rightarrow \mathbb{R}_+$ with the following properties: (1) Identifiability $\rho(x, y) = 0 \Leftrightarrow x = y$, (2) Symmetry $\rho(x, y) = \rho(y, x)$, (3) Sub-additivity $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$.*

Example: $\mathbb{R}^d$ with the Euclidean distance $\rho(x, y) = \sqrt{\sum_{i=1}^d (x_i - y_i)^2}$ is a metric. This is the only one we will use in this lecture, but covering number bounds apply much more generally.

Definition 3 (Covering number)**.** *Let (S, ρ) be a metric space, and let $T \subset S$. We say that $T' \subset S$ is an α-**cover** for T if, for all $x \in T$, there exists $y \in T'$ such that $\rho(x, y) \leq \alpha$. The α-**covering number** of (T, ρ), denoted $\mathcal{N}(\alpha, T, \rho)$ is the size of the smallest α-covering. The **metric entropy** is the log covering number.*

We will see some examples shortly, but let us first derive a simple bound on the rademacher complexity in terms of the metric entropy.

Proposition 4. *For $A \subset \mathbb{R}^n$,*

$$\mathcal{R}(A) \leq \inf_{\alpha>0} \left\{ \max_a \|a\|_2 \frac{\sqrt{2\log(\mathcal{N}(\sqrt{n}\alpha, A, \ell_2))}}{n} + \alpha \right\}.$$

Proof. Fix $\alpha > 0$ and let $\tilde{A}$ be an $\sqrt{n}\alpha$-cover of A in ℓ_2 of size $|\tilde{A}| = \mathcal{N}(\sqrt{n}\alpha, A, \ell_2)$. Note that one is guaranteed to exist by the definition of $\mathcal{N}$. For element a, let $\tilde{a}$ be the covering element, Then,

$$\mathcal{R}(A) = \mathbb{E}_\sigma \sup_a \frac{1}{n}\langle \sigma, a\rangle = \mathbb{E}_\sigma \sup_a \frac{1}{n}\langle \sigma, \tilde{a}\rangle + \frac{1}{n}\langle \sigma, a - \tilde{a}\rangle$$

$$\mathbb{E}_\sigma \sup_a \frac{1}{n}\langle \sigma, \tilde{a}\rangle + \frac{1}{n}\|\sigma\|_2\|a - \tilde{a}\|_2 \leq \mathcal{R}(\tilde{A}) + \alpha \leq \max_a \|a\|_2 \frac{\sqrt{2\log(\mathcal{N}(\sqrt{n}\alpha, A, \ell_2))}}{n} + \alpha$$

This argument applies for any α, so the result follows by taking the minimal α. □

Note that a similar bound can also be proved using $\mathcal{N}(n\alpha, A, \ell_1)$ by applying Holder's inequality instead of Cauchy-Schwarz in the proof.

Proposition 5.

$$\mathcal{R}(A) \leq \inf_{\alpha>0} \left\{ \max_a \|a\|_2 \frac{\sqrt{2\log(\mathcal{N}(n\alpha, A, \ell_1))}}{n} + \alpha \right\}.$$

Remark 6. *Often the covering numbers are expressed in a slightly different way, where the ℓ_2 metric is instead defined as $\ell_{2,n}(x, y) = \sqrt{1/n\sum_{i=1}^n (x_i - y_i)^2}$. In this case the bound depends on* $\log(\mathcal{N}(\alpha, A, \ell_{2,n}))$.

2.1 Examples

Linear functions. Let $A \subset \mathbb{R}^n$, with $c = \max_a \|a\|_2$ and assume that A lies in a d dimensional subspace of $\mathbb{R}^n$. After use contraction to eliminate the loss, this is precisely the setting of linear regression from last time, since $A = \{(\langle w, x_1\rangle, \ldots, \langle w, x_n\rangle)\}_{w \in \mathcal{W}_2}$, where $\mathcal{W}_2$ is the unit Euclidean ball. So if $X \in \mathbb{R}^{n\times d}$ then A is contained in the column space of X which has dimension at most d.

We show that $\mathcal{N}(\alpha, A, \ell_2) \leq (2c\sqrt{d}/\alpha)^d$. To see why, let $v_1, \ldots, v_d$ be an orthonormal basis for span(A), which means we can write any vector $a \in A$ as $\sum_{j=1}^d \alpha_j v_j$ where $\|\alpha\|_\infty \leq \|\alpha\|_2 \leq \|a\|_2 \leq c$. Now consider the set

$$\tilde{A} = \left\{ \sum_{i=1}^d \tilde{\alpha}_i v_i : \forall i, \tilde{\alpha}_i \in \{-c, -c + \alpha/\sqrt{d}, -c + 2\alpha/\sqrt{d}, \ldots, c\} \right\}$$

For any a, we can find $\tilde{a}$ by rounding the coefficients α_i for a such that

$$\|a - \tilde{a}\|^2 = \|\sum_{i=1}^d (\alpha_i - \tilde{\alpha}_i)v_i\|_2^2 = \sum_{i=1}^d (\alpha_i - \tilde{\alpha}_i)^2 \leq \alpha^2$$

which means that $\tilde{A}$ is an α-cover. The number of elements is $(2c\sqrt{d}/\alpha)^d$.

Without instantiating c, this gives us the bound

$$\mathcal{R}(A) \leq \inf_\alpha c\frac{\sqrt{2d\log(c\sqrt{d}/(\sqrt{n}\alpha))}}{n} + \alpha,$$

Now if $c = O(\sqrt{n})$, which is what we expect if $a \in [-1, 1]^n$ then we set $\alpha = \sqrt{d/n}$ and get $O(\sqrt{d\log(n)/n})$.

For linear regression it seems plausible that actually $c = \sqrt{n/d}$ in benign cases. If the distribution is somewhat spherical, then we would expect $\|X\|_2 \leq O(\sqrt{n/d})$. To see why, think about if $x_i \sim \text{Unif}(\{e_i\}_{i=1}^d)$ then since $X \in \mathbb{R}^{n\times d}$, $\|X\| = \sqrt{\lambda_{\max}(XX^T)}$ which is a $d \times d$ diagonal matrix and we expect each diagonal to be roughly n/d. Making this more precise would lead to an $O(\sqrt{\log(n)/n})$ bound. This is closer to what we saw in last lecture when we worked with the Rademacher complexity directly, but really covering numbers are most useful for nonparametric classes.

Remark 7. *There is actually a better proof that gives a $(2c/\alpha+1)^d = O((c/\alpha)^d)$ bound for covering of the d-dimensional ball of radius c, in the same norm. The proof is more volumetric but the bound is better in that there is no $\sqrt{d}$ dependence. The proof idea is based on introducing a packing, which is a set of points such that no two points are within α of each other. It is not too hard to see that covering number is at most the size of the maximal packing, since if you cannot pack another point in at distance α, then all points must be covered at scale α. Then take the maximal α packing for the ball $\{x_1, \ldots, x_M\}$ and observe that the balls $B(x_i, \alpha/2)$ are disjoint and if you take $\bigcup B(x_i, \alpha/2)$ then this set is contained in $B(0, c+\alpha/2)$. Using that the radius and volume are related by exponentiating by d you'll get the result.*

All functions. A special case of the above is, if $A = [-1,1]^n$ is the set of *all vectors* then the covering number is $\mathcal{N}(\alpha, A, \ell_2) \sim (n/\alpha)^n$. Then since in applying Massart's lemma $\max_a \|a\| = \sqrt{n}$, we obtain a vacuous bound $\mathcal{R}(A) \le O(1)$, which is not very useful.

Monotonic functions. Let $A \subset [-1,1]^n$ be the monotonic vectors so that $a_i \le a_{i+1} \ldots$. We prove that $\mathcal{N}(\sqrt{n}\alpha, A, \ell_2) = O(n^{2/\alpha})$ which means that $\mathcal{N}(\alpha, A, \ell_2) = O(n^{\sqrt{n}/\alpha})$. To see the first point, discretize $[-1,1]$ to $2/\alpha$ levels of size α. The cover will be formed by choosing for each level, an index $i \in [n]$ at which the vector increases above that level. Clearly then this cover has size at most $n^{2/\alpha}$, but moreover, for each $a \in A$, we know that there is a cover element that is at every point at most α away from this function, so the ℓ_2 norm is at most $\sqrt{n}\alpha$. As before $\max_a \|a\|_2 = \sqrt{n}$ so our discretization bound is

$$\mathcal{R}(A) \le \inf_{\alpha>0} \sqrt{n} \times \frac{\sqrt{2/\alpha \log(n)}}{n} + \alpha = \inf_{\alpha>0} \sqrt{\frac{2\log(n)}{n\alpha}} + \alpha = \tilde{O}\left(\left(\frac{\log n}{n}\right)^{-1/3}\right)$$

3 Chaining bounds

In the proof, we chose a scale α to minimize the discretization error and the finite-class rademacher complexity. While this can work well at times, it can often be better to, in some sense choose all scales simultaneously, or look at increasing refinements of the metric space.

Theorem 8.

$$\mathcal{R}(A) \le 12 \int_0^\infty \frac{\sqrt{\log(\mathcal{N}(\alpha, A, \ell_2))}}{n} d\alpha$$

Remark 9. *It is probably more common to see this stated using the $\ell_{2,n}$ metric, in which case you would have a $1/\sqrt{n}$ dependence. However this translation is accounted for here since at some point $\mathcal{N}(\alpha, A, \ell_2) = 1$ in which case the integrand is zero. This depends on $B = \max_a \|a\|_2$ which is $\sqrt{n}$ larger in the ℓ_2 metric than in the $\ell_{2,n}$ metric.*

Proof. Let $B = \max_a \|a\|_2$. We will pick the scales $\alpha_0 = B, \ldots \alpha_i = 2^{-i}B$, and we let T_i be an α_i cover of A in ℓ_2. Fixing a, let $\hat{a}^{(i)}$ be the covering element for a at scale i, so that $\|a - \hat{a}^{(i)}\|_2 \le \alpha_i$ and without let $T_0 = \{0\} = \{\hat{a}^{(0)}\}$. Since for every a and any N we may write

$$a = a + \sum_{j=1}^{N} (\hat{a}^{(j)} - \hat{a}^{(j-1)}) + \hat{a}^{(0)} - \hat{a}^{(N)}$$

we can use this representation in the Rademacher average definition

$$\begin{aligned}
\mathcal{R}(A) &= \mathbb{E} \sup_a \frac{1}{n} \sum_{i=1}^n \sigma_i a_i = \mathbb{E} \sup_a \left\{ \frac{1}{n} \sum_{i=1}^n \sigma_i (a_i - \hat{a}_i^{(N)}) + \frac{1}{n} \sum_{i=1}^n \sigma_i \left(\sum_{j=1}^N \hat{a}_i^{(j)} - \hat{a}_i^{(j-1)} \right) \right\} \\
&= \mathbb{E} \sup_a \frac{1}{n} \langle \sigma, a - \hat{a}^{(N)} \rangle + \sum_{j=1}^N \mathbb{E} \sup_a \frac{1}{n} \langle \sigma, a^{(j)} - a^{(j-1)} \rangle
\end{aligned}$$

For the first term, we will simply use that $\langle \sigma, a - \hat{a}^{(N)} \rangle \leq \|\sigma\| \|a - \hat{a}^{(N)}\| \leq \sqrt{n}\alpha_N$. For the second term we'll do something similar, but we can see that $\|a^{(j)} - a^{(j-1)}\| \leq \alpha_j + \alpha_{j-1} \leq 3\alpha_j$, since $\alpha_{j-1} = 2\alpha_j$. Thus applying Massart's lemma to these latter terms, we get

$$\mathcal{R}(A) \leq \frac{\alpha_N}{\sqrt{n}} + \sum_{j=1}^{N} \frac{3\alpha_j}{n} \sqrt{2 \log(|T_j| |T_{j-1}|)} \leq \frac{\alpha_N}{\sqrt{n}} + \frac{6}{n} \sum_{j=1}^{N} \alpha_j \sqrt{\log(|T_j|)}$$

Some versions of the bound are stated this way and this seems fine. To obtain the integral, we look at $\sqrt{n}$ times the right hand side and use the fact that $\alpha_j = 2(\alpha_j - \alpha_{j+1})$.

$$\begin{aligned} \alpha_N + 12 \sum_{j=1}^{N} (\alpha_j - \alpha_{j+1}) \sqrt{\frac{\log(|T_j|)}{n}} &\leq \alpha_N + 12 \sum_{j=1}^{N} \int_{\alpha_{j+1}}^{\alpha_j} \sqrt{\frac{\log(\mathcal{N}(\alpha, A, \ell_2))}{n}} d\alpha \\ &\leq \alpha_N + 12 \int_{\alpha_{N+1}}^{\alpha_0} \sqrt{\frac{\log(\mathcal{N}(\alpha, A, \ell_2))}{n}} d\alpha \end{aligned}$$

Now push $N \to \infty$ so that $\alpha_N \to 0$. □

Linear class. With $c = \sqrt{n}$ we know that $\mathcal{N}(\alpha, A, \ell_2) \leq (2\sqrt{n}/\alpha)^d$ and if $\alpha > \sqrt{n}$ we know that the covering number is 1, so the log covering number is 0 and we can update the limits of integration to get

$$\frac{12\sqrt{d}}{n} \int_0^{\sqrt{n}} \sqrt{\log(2\sqrt{n}/\alpha)} d\alpha$$

By applying a change of variables $\alpha = 2\sqrt{n} \exp(-y^2)$ this terms looks like the variance of a gaussian.

$$\int_0^{\sqrt{n}} \sqrt{\log(2\sqrt{n}/\alpha)} d\alpha \leq \int_0^{\infty} 4\sqrt{n} y^2 \exp(-y^2) dy = 4\sqrt{2n/\pi}.$$

So the entire bound is $O(\sqrt{d/n})$, which removed the $\log(n)$ factor from the discretization bound.

Monotonic class. For monotonic functions, we saw that $\mathcal{N}(\alpha, A, \ell_2) \leq n^{2\sqrt{n}/\alpha}$, and as before we only have to integrate up to $\sqrt{n}$. Applying the theorem gives

$$\mathcal{R}(A) \leq \frac{12}{n} \int_0^{\sqrt{n}} \sqrt{\frac{2\sqrt{n}\log(n)}{\alpha}} = \frac{12\sqrt{2\log n}}{n^{3/4}} \int_0^{\sqrt{n}} \sqrt{\frac{1}{\alpha}} d\alpha = \frac{6\sqrt{2\log n}}{n^{3/4}} \left(\sqrt{\alpha} \, |_0^{\sqrt{n}} \right) = O(\sqrt{\log n / n}).$$

which is much faster than the $n^{-1/3}$ rate that we saw before.

4 Recap

That's it for empirical process theory. To summarize:

1. We saw many ways to establish uniform convergence, including in terms of VC-dimension, Rademacher complexity, and Covering number bounds.
2. We also saw many useful proof strategies, including symmetrization, discretization, the log-sum-exp method, chernoff method and several others.
3. Lastly we saw how to apply this technology to obtain excess risk or sample complexity bounds for a number of learning problems.
4. I should also point out that there are many other ways to prove that a learning algorithm will generalize well, beyond uniform convergence. For example, if a learning algorithm is *stable* to small perturbations of the training data, it cannot overfit too much so it will generalize. There are also more advanced empirical process notions like localization that are beyond the scope of this course.

We could dedicate an entire course to empirical process theory, but now we are going to move on to some learning algorithms. Next time, we'll see how to use some of this technology in regression and classification settings.