18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet
Scribe: Zach Izzo

Lecture 7
Sep. 30, 2015

In this lecture, we continue our discussion of covering numbers and compute upper bounds for specific conditional Rademacher averages $\hat{\mathcal{R}}_n^x(\mathcal{F})$. We then discuss chaining and conclude by applying it to learning.

Recall the following definitions. We define the risk function

$$R(f) = \mathbb{E}[\ell(X, f(X))], \qquad (X,Y) \in \mathcal{X} \times [-1,1]\,,$$

for some loss function $\ell(\cdot,\cdot)$. The conditiona Rademacher average that we need to control is

$$\mathcal{R}(\ell l \circ \mathcal{F}) = \sup_{(x_1,y_1),\ldots,(x_n,y_n)} \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i \ell(y_i, f(x_i))\right|\right].$$

Furthermore, we defined the conditional Rademacher average for a point $x = (x_1,\ldots,x_n)$ to be

$$\hat{\mathcal{R}}_n^x(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i)\right|\right].$$

Lastly, we define the ε-covering number $N(\mathcal{F}, d, \varepsilon)$ to be the minimum number of balls (with respect to the metric d) of radius ε needed to cover $\mathcal{F}$. We proved the following theorem:

Theorem: Assume $|f| \le 1$ for all $f \in \mathcal{F}$. Then

$$\hat{\mathcal{R}}_n^x(\mathcal{F}) \le \inf_{\varepsilon>0}\left\{\varepsilon + \sqrt{\frac{2\log(2N(\mathcal{F}, d_1^x, \varepsilon))}{n}}\right\},$$

where d_1^x is given by

$$d_1^x(f,g) = \frac{1}{n}\sum_{i=1}^n |f(x_i) - g(x_i)|.$$

We make use of this theorem in the following example. Define $B_p^d = \{x \in \mathbb{R}^d : |x|_p \le 1\}$. Then take $f(x) = \langle a, x\rangle$, set $\mathcal{F} = \{\langle a, \cdot\rangle : a \in B_\infty^d\}$, and $\mathcal{X} = B_1^d$. By Hölder's inequality, we have

$$|f(x)| \le |a|_\infty |x|_1 \le 1,$$

so the theorem above holds. We need to compute the covering number $N(\mathcal{F}, d_1^x, \varepsilon)$. Note that for all $a \in B_\infty^d$, there exists $v = (v_1,\ldots,v_n)$ such that $v_i = g(x_i)$ and

$$\frac{1}{n}\sum_{i=1}^n |\langle a, x_i\rangle - v_i| \le \varepsilon$$

for some function g. For this case, we will take $g(x) = \langle b, x\rangle$, so $v_i = \langle b, x_i\rangle$. Now, note the following. Given this definition of g, we have

$$d_1^x(f,g) = \frac{1}{n}\sum_{i=1}^n |\langle a, x_1\rangle - \langle b, x_i\rangle| = \frac{1}{n}\sum_{i=1}^n |\langle a-b, x_i\rangle| \le |a-b|_\infty$$

by Hölder's inequality and the fact that $|x|_1 = 1$. So if $|a-b|_\infty \le \varepsilon$, we can take $v_i = \langle b, x_i \rangle$. We just need to find a set of $\{b_1, \ldots, b_M\} \subset \mathbb{R}^d$ such that, for any a there exists b_j such that $|a - b_j|_\infty < \infty$. We can do this by dividing B_∞^d into cubes with side length ε and taking the b_j's to be the set of vertices of these cubes. Then any $a \in B_\infty^d$ must land in one of these cubes, so $|a - b_j|_\infty \le \varepsilon$ as desired. There are c/ε^d of such b_j's for some constant $c > 0$. Thus

$$N(B_\infty^d, d_1^x, \varepsilon) \le c/\varepsilon^d.$$

We now plug this value into the theorem to obtain

$$\hat{\mathcal{R}}_n^x(\mathcal{F}) \le \inf_{\varepsilon \ge 0} \left\{ \varepsilon + \sqrt{\frac{2\log(c/\varepsilon^d)}{n}} \right\}.$$

Optimizing over all choices of ε gives

$$\varepsilon^* = c\sqrt{\frac{d\log(n)}{n}} \qquad \Rightarrow \qquad \hat{\mathcal{R}}_n^x(\mathcal{F}) \le c\sqrt{\frac{d\log(n)}{n}}.$$

Note that in this final inequality, the conditional empirical risk no longer depends on x, since we "sup'd" x out of the bound during our computations. In general, one should ignore x unless it has properties which will guarantee a bound which is better than the sup. Another important thing to note is that we are only considering one granularity of $\mathcal{F}$ in our final result, namely the one associated to ε^*. It is for this reason that we pick up an extra log factor in our risk bound. In order to remove this term, we will need to use a technique called *chaining*.

5.4 Chaining

We have the following theorem.

Theorem: Assume that $|f| \le 1$ for all $f \in \mathcal{F}$. Then

$$\hat{\mathcal{R}}_n^x \le \inf_{\varepsilon > 0} \left\{ 4\varepsilon + \frac{12}{\sqrt{n}} \int_\varepsilon^1 \sqrt{\log(N(\mathcal{F}, d_2^x, t))} dt \right\}.$$

(Note that the integrand decays with t.)

Proof. Fix $x = (x_1, \ldots, x_n)$, and for all $j = 1, \ldots, N$, let V_j be a minimal 2^{-j}-net of $\mathcal{F}$ under the d_2^x metric. (The number N will be determined later.) For a fixed $f \in \mathcal{F}$, this process will give us a "chain" of points f_i° which converges to f: $d_2^x(f_i^\circ, f) \le 2^{-j}$.

Define $F = \{(f(x_1), \ldots, f(x_n))^\top,\ f \in \mathcal{F}\} \subset [-1,1]^n$. Note that

$$\hat{\mathcal{R}}_n^x(\mathcal{F}) = \frac{1}{n}\mathbb{E}\sup_{f \in F}\langle \sigma, f \rangle$$

where $\sigma = (\sigma_1, \ldots, \sigma_n)$. Observe that for all N, we can rewrite $\langle \sigma, f \rangle$ as a telescoping sum:

$$\langle \sigma, f \rangle = \langle \sigma, f - f_N^\circ \rangle + \langle \sigma, f_N^\circ - f_{N-1}^\circ \rangle + \ldots + \langle \sigma, f_1^\circ - f_0^\circ \rangle$$

where $f_0^\circ := 0$. Thus

$$\hat{\mathcal{R}}_n^x(\mathcal{F}) \le \frac{1}{n}\mathbb{E}\sup_{f\in F}|\langle \sigma, f - f_N^\circ\rangle| + \sum_{j=1}^{N}\frac{1}{n}\mathbb{E}\sup_{f\in F}|\langle \sigma, f_j^\circ - f_{j-1}^\circ\rangle|\,.$$

We can control the two terms in this inequality separately. Note first that by the Cauchy-Schwarz inequality,

$$\frac{1}{n}\mathbb{E}\sup_{f\in F}|\langle \sigma, f - f_N^\circ\rangle| \le |\sigma|_2\frac{d_2^x(f, f_N^\circ)}{\sqrt{n}}.$$

Since $|\sigma|_2 = \sqrt{n}$ and $d_2^x(f, f_N^\circ) \le 2^{-N}$, we have

$$\frac{1}{n}\mathbb{E}\sup_{f\in F}|\langle \sigma, f - f_N^\circ\rangle| \le 2^{-N}\,.$$

Now we turn our attention to the second term in the inequality, that is

$$S = \sum_{j=1}^{N}\frac{1}{n}\mathbb{E}\sup_{f\in F}|\langle \sigma, f_j^\circ - f_{j-1}^\circ\rangle|.$$

Note that since $f_j^\circ \in V_j$ and $f_{j-1}^\circ \in V_{j-1}$, there are at most $|V_j||V_{j-1}|$ possible differences $f_j^\circ - f_{j-1}^\circ$. Since $|V_{j-1}| \le |V_j|/2$, $|V_j||V_{j-1}| \le |V_j|^2/2$ and we find ourselves in the finite dictionary case. We employ a risk bound from earlier in the course to obtain the inequality

$$\mathcal{R}_n(B) \le \max_{b\in B}|b|_2\frac{\sqrt{2\log(2|B|)}}{n}\,.$$

In the present case, $B = \{f_j^\circ - f_{j-1}^\circ\,, f \in F\}$ so that $|B| \le |V_j|^2/2$. It yields

$$\mathcal{R}_n(B) \le r\cdot\frac{\sqrt{2\log(\frac{2|V_j|^2}{2})}}{n} = 2r\cdot\frac{\sqrt{\log|V_j|}}{n}\,,$$

where $r = \sup_{f\in F}|f_j^\circ - f_{j-1}^\circ|_2$. Next, observe that

$$|f_j^\circ - f_{j-1}^\circ|_2 = \sqrt{n}\cdot d_2^x(f_j^\circ, f_{j-1}^\circ) \le \sqrt{n}(d_2^x(f_j^\circ, f) + d_2^x(f, f_{j-1}^\circ)) \le 3\cdot 2^{-j}\sqrt{n}\,.$$

by the triangle inequality and the fact that $d_2^x(f_j^\circ, f) \le 2^{-j}$. Substituting this back into our bound for $\mathcal{R}_n(B)$, we have

$$\mathcal{R}_n(B) \le 6\cdot 2^{-j}\sqrt{\frac{\log|V_j||}{n}} = 6\cdot 2^{-j}\sqrt{\frac{\log(N(\mathcal{F}, d_2^x, 2^{-j}))}{n}}$$

since V_j was chosen to be a minimal 2^{-j}-net.

The proof is almost complete. Note that $2^{-j} = 2(2^{-j} - 2^{-j-1})$ so that

$$\frac{6}{\sqrt{n}}\sum_{j=1}^{N}2^{-j}\sqrt{\log(N(\mathcal{F}, d_2^x, 2^{-j}))} = \frac{12}{\sqrt{n}}\sum_{j=1}^{N}(2^{-j} - 2^{-j-1})\sqrt{\log(N(\mathcal{F}, d_2^x, 2^{-j}))}\,.$$

Next, by comparing sums and integrals (Figure 1), we see that

$$\sum_{j=1}^{N}(2^{-j} - 2^{-j-1})\sqrt{\log(N(\mathcal{F}, d_2^x, 2^{-j}))} \le \int_{2^{-(N+1)}}^{1/2}\sqrt{\log(N(\mathcal{F}, d_2^x, t))}dt.$$

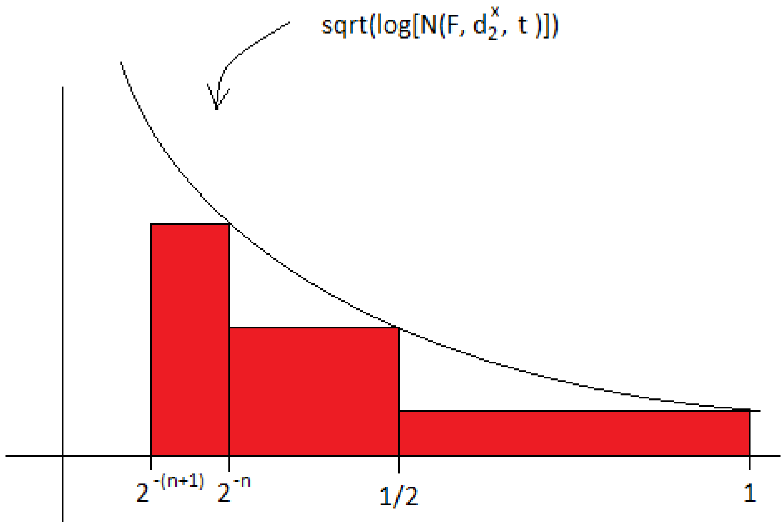

Figure 1: A comparison of the sum and integral in question.

So we choose N such that $2^{-(N+2)} \leq \varepsilon \leq 2^{-(N+1)}$, and by combining our bounds we obtain

$$\hat{\mathcal{R}}_n^x(\mathcal{F}) \leq 2^{-N} + \frac{12}{\sqrt{n}} \int_{2^{-(N+1)}}^{1/2} \sqrt{\log(N(\mathcal{F}, d_2^x, t))} dt \leq 4\varepsilon + \int_{\varepsilon}^{1} \sqrt{\log(N, \mathcal{F}, t)} dt$$

since the integrand is non-negative. (Note: this integral is known as the "Dudley Entropy Integral.") □

Returning to our earlier example, since $N(\mathcal{F}, d_2^x, \varepsilon) \leq c/\varepsilon^d$, we have

$$\hat{\mathcal{R}}_n^x(\mathcal{F}) \leq \inf_{\varepsilon > 0} \left\{ 4\varepsilon + \frac{12}{\sqrt{n}} \int_{\varepsilon}^{1} \sqrt{\log((c'/t)^d)} dt \right\}.$$

Since $\int_0^1 \sqrt{\log(c/t)} dt = \bar{c}$ is finite, we then have

$$\hat{\mathcal{R}}_n^x(\mathcal{F}) \leq 12\bar{c}\sqrt{d/n}.$$

Using chaining, we've been able to remove the log factor!

5.5 Back to Learning

We want to bound

$$\mathcal{R}_n(\ell \circ \mathcal{F}) = \sup_{(x_1,y_1),\ldots,(x_n,y_n)} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(y_i, f(x_i)) \right| \right].$$

We consider $\hat{\mathcal{R}}_n^x(\Phi \circ \mathcal{F}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \Phi \circ f(x_i) \right| \right]$ for some L-Lipschitz function Φ, that is $|\Phi(a) - \Phi(b)| \leq L|a - b|$ for all $a, b \in [-1, 1]$. We have the following lemma.

Theorem: (Contraction Inequality) Let Φ be L-Lipschitz and such that $\Phi(0) = 0$, then

$$\hat{\mathcal{R}}_n^x(\Phi \circ \mathcal{F}) \leq 2L \cdot \hat{\mathcal{R}}_n^x(\mathcal{F})\,.$$

The proof is omitted and the interested reader should take a look at [LT91, Kol11] for example.

As a final remark, note that requiring the loss function to be Lipschitz prohibits the use of $\mathbb{R}$-valued loss functions, for example $\ell(Y, \cdot) = (Y - \cdot)^2$.

References

[Kol11] Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems. École d'Été de Probabilités de Saint-Flour XXXVIII-2008.* Lecture Notes in Mathematics 2033. Berlin: Springer. ix, 254 p. EUR 48.10 , 2011.

[LT91] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)].* Springer-Verlag, Berlin, 1991. Isoperimetry and processes.