EECS 281B / STAT 241B: Advanced Topics in Statistical Learning **Spring 2009**

Lecture 20 — April 6th

Lecturer: Martin Wainwright *Scribe: Vladislav Voroninski*

Note: These lecture notes are still rough, and have only have been mildly proofread.

This is the danger environment.

Outline

- Rademacher and empirical covering
- some model selection issues

20.1 Recap

In recent lectures, we have talked we talked about Rademacher complexity:

$$\widehat{\mathbb{R}_n}(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma^i f(X^{(i)}) \right| \right] \tag{20.1}$$

We have seen the connection to shatter coefficients and VC dimension. Today, we discuss connections to *empirical covering numbers*. Recall the definition of the mpirical L_1-norm, denoted by $L_1(\widehat{\mathbb{P}_n})$ and defined by

$$||f - g||_{L_1(\widehat{\mathbb{P}_n})} = \frac{1}{n} \sum_{i=1}^{n} \left| f(X^{(i)}) - g(X^{(i)}) \right|. \tag{20.2}$$

Similarly, we define the empirical L_2-norm

$$||f - g||_{L_2(\widehat{\mathbb{P}}_n)} = \frac{1}{n} \sqrt{\sum_{i=1}^{n} (f(X^{(i)}) - g(X^{(i)}))^2} \tag{20.3}$$

We refer to metric entropies based on these norms as empirical metric entropies.

20.2 Rademacher complexity and empirical metric entropy

Our first result provides a connection between the Rademacher complexity and the empirical metric entropy:

Theorem 20.1 (Discretization). *The Rademacher complexity is upper bounded as*

$$\widehat{\mathbb{R}}_n(\mathcal{F}) \leq \inf_{t>0}\left\{ t + c\sqrt{\frac{\log N(t, \mathcal{F}, L_2(\widehat{\mathbb{P}}_n))}{n}} \right\}, \tag{20.4}$$

where c is some constant and the square-root term is the empirical $L_2(\widehat{\mathbb{P}}_n)$ metric entropy.

Proof: Let $N = N(t, \mathcal{F}, L_2(\widehat{\mathbb{P}}_n))$ and let $f_1, \ldots, f_N$ be a t-cover of $\mathcal{F}$ w.r.t the $L_2(\widehat{\mathbb{P}_n})$ norm. Then for any $f \in \mathcal{F}$, there exists a function f_k such that $||f_k - f||_{L_2(\widehat{\mathbb{P}_n})} \leq t$. Thus, we have

$$\begin{aligned} \left|\frac{1}{n}\sum_{i=1}^{n} \sigma^{(i)} f(X^{(i)})\right| &\leq \left|\frac{1}{n}\sum_{i=1}^{n} \sigma^{(i)} f_k(X^{(i)})\right| + \left|\frac{1}{n}\sum_{i=1}^{n} \sigma^{(i)} (f(X^{(i)}) - f_k(X^{(i)}))\right| \\ &\leq \left|\frac{1}{n}\sum_{i=1}^{n} \sigma^{(i)} f(X^{(i)})\right| + \frac{1}{n}||\sigma||_{L_2(\widehat{\mathbb{P}_n})}||f - f_k||_{L_2(\widehat{\mathbb{P}_n})}, \end{aligned}$$

where the second step applies the Cauchy-Schwarz inequality to the inner product defined by $\langle f,\, g\rangle := \frac{1}{n}\sum_{i=1}^{n} f(X^{(i)})g(X^{(i)})$.

Thus, we conclude that

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} \sigma^{(i)} f(X^{(i)})\right| \leq \max_{k=1,2\ldots n}\left|\frac{1}{n}\sum_{i=1}^{n} \sigma^{(i)} f_k(X^{(i)})\right| + t,$$

and hence,

$$\begin{aligned} \widehat{\mathbb{R}_n}(\mathcal{F}) &\leq \mathbb{E}_\sigma \max_{k=1,2\ldots n}\left|\frac{1}{n}\sum_{i=1}^{n} \sigma^{(i)} f_k(X^{(i)})\right| + t \\ &\leq \sqrt{\frac{\log|N|}{n}} + t, \end{aligned}$$

where the second inequality uses the form of Rademacher complexity for the finite set of functions $\mathcal{F}_t$. Finally, we can take the infimum over $t > 0$ on the RHS to obtain the result. □

Let's consider some examples to illustrate.

Example 1. *We begin by considering the case of Lipschitz functions that map $[0,1]^d$ to the real line $\mathbb{R}$. It is known that $\log N(t, \mathcal{F}, ||.||)\ (\frac{1}{t})^d$ as $t \to 0$. In this case, we have*

$$\widehat{\mathbb{R}_n}(\mathcal{F}) \leq \inf_{t>0}\{t + c\sqrt{\frac{1}{n}(\frac{1}{t})^d}\} \leq c(\frac{1}{n})^{\frac{1}{2+d}}, \tag{20.5}$$

obtained by choosing $t = (\frac{1}{n})^{\frac{1}{2+d}}$. This very slow decay of the Rademacher complexity for large dimension d is the usual manifestation of the curse of dimensionality: i.e., to reduce the Rademacher complexity to some level ϵ, we need roughly $n \asymp (1/\delta)^{2+d}$ samples, which explodes exponentially in the dimension.

We can obtain faster rates of decay of the Rademacher complexity by imposing more structure on our functions.

Example 2. *As an extension of the previous example, let us consider functions on* $[0,1]^d$ *with* k *derivatives, and assume that the* k*th derivative is a Lipschitz function. In this case, it is known that*

$$\log N(t, \mathcal{F}, ||.||) \;\; (\frac{1}{t})^{\frac{d}{k+1}}$$

so that we can conclude that

$$\widehat{\mathbb{R}}_n(\mathcal{F}_k) \leq c(\frac{1}{n})^{\frac{k+1}{2(k+1)+d}}. \tag{20.6}$$

If d *remains fixed while the number of derivatives* k *increases, then we obtain faster and faster rates, which is intuitively reasonable. However, if* d *is thought of as very large (or even increasing), then we also need the degree of smoothness* k *to be very large (or increasing) if there is any hope of obtaining a reasonable rate of convergence.*

In the previous examples (unless the degree of smoothness is very large), the rates will be very slow for high dimensions d; note that even $d = 100$ and $\delta = 0.1$ in Example 1 would yield a required sample size of order $n \asymp 10^{102}$, which is more than the number of atoms in the universe. There are other ways of side-stepping the curse of dimensionality, which involve imposing additive or separable structure on spaces of functions.

Example 3. *Given a function* $f : [0,1]^d \to \mathbb{R}$*, suppose that we assume it obeys an additive decomposition, of the form*

$$f(x_1, x_2 \ldots x_d) = \sum_{i=1}^{d} g_i(x_i) \tag{20.7}$$

where each $g_i : \mathbb{R} \to \mathbb{R}$*. In this case, we would expect to have faster rates, since the function is not really* d*-variate in full generality, but rather a sum of univariate functions.*

A related model is a sparsity model in which we assume that the function f *depends only on some subset* $S \subset \{1, 2 \ldots d\}, |S| = k < d$ *of co-ordinates—say*

$$f(x_1, x_2 \ldots x_d) = g(x_S) \tag{20.8}$$

where $g : \mathbb{R}^k \to \mathbb{R}$*. In this case, if the subset* S *were known, then we could immediately restrict to these co-ordinates. Otherwise, one can imagine trying to estimate the subset* S*, and then performing regression on the restricted subset.*

Our final example concerns a parametric class of functions, namely the class of all linear funcitons on the unit ball

Example 4. *Let us consider the class of linear functions*

$$\mathcal{F} = \{f_\theta(x) =< \theta, x > | \theta \in \mathbb{R}^d, ||\theta||_2 = 1\}, \tag{20.9}$$

which is parameterized by vectors on the unit ball. In this case, it can be shown (see Homework # 3) that

$$\log N(t, \mathcal{F}, ||.||_2)\ d(\frac{1}{t}), \qquad \textit{valid for } t \to 0. \tag{20.10}$$

Using this relationship in the discretization theorem, we obtain

$$\widehat{\mathbb{R}}_n(\mathcal{F}) \leq \inf_{t>0} \left\{ t + c_1 \sqrt{\frac{d \log(\frac{1}{t})}{n}} \right\} \leq c_2 \sqrt{\frac{d \log n}{n}} \tag{20.11}$$

The scaling $\sqrt{d/n}$ in this upper bound is of the right order, but the $\log n$ term is actually an artifact of the method.

20.3 Dudley's entropy integral

The material that we have been covering is closely related to empirical process theory, a branch of probability and statistics that is concerned with the behavior of stochastic processes that are indexed by functions or other objects. A key result in this area is Dudley's entropy integral, which provides a much sharper upper bound on Rademacher complexity than the simple discretization argument that we considered. We state the result here:

Theorem 20.2. *There exists a constant c such that the Rademacher complexity is upper bounded as:*

$$\widehat{\mathbb{R}}_n(\mathcal{F}) \leq c \int_0^\infty \sqrt{\frac{\log N(t, \mathcal{F}, L_2(\widehat{\mathbb{P}}_n))}{n}} dt. \tag{20.12}$$

In the integral on the RHS, the important part is its behavior as $t \to 0$. (If the function class has a finite diameter D, then we known that the covering number is 1 for t sufficiently large, so the upper integration limit can be made finite.)

This theorem can be proved via the *chaining argument*, in which we decompose the supremum defining the Rademacher (or Gaussian) complexity into a series of terms, and discretize each term in a refined manner.

To illustrate the consequences of Dudley's theorem, let us revisit the case of the unit ball in d dimensions. Using our previous discretization method, the best bound on the Rademacher complexity scaled as $\sqrt{\frac{d \log n}{n}}$. This can be sharpened via the Dudley integral:

Example 5. *In the case of the unit ℓ_2 ball $B_2(0,1)$, from Dudley's theorem and known results on the metric entropy of the ℓ_2 ball, we have*

$$\begin{aligned} \widehat{\mathbb{R}}_n(B(0,1)) &\leq c_1 \int_0^D \sqrt{\frac{\log N(t, B(0,1), ||.||)}{n}} dt \\ &= c_1 \sqrt{\frac{d}{n}} \int_0^1 \sqrt{\log(\frac{1}{t})} dt, \end{aligned}$$

where we have used the fact that the ℓ_2 *ball has diameter* 1*. Continuing on, it can be shown that the given integral is finite, so that we conclude that*

$$\widehat{\mathbb{R}}_n(B(0,1)) \;\; \leq \;\; c_2\sqrt{\frac{d}{n}}.$$

Note that we have removed the superfluous $\log n$ *factor from the previous result.*