9 Chaining and Dudley's Theorem

In this lecture, we revisit some of the techniques we covered in Section 5, but there is one key difference. Before, we used the ϵ-nets to discretize the uncountable function class to be able to apply union bound and obtain a generalization bound. In this section, we will use this technique to bound the Rademacher complexity of the function class which in turn will imply generalization.

9.1 ϵ-Nets revisited

Previously in Section 5, we covered the parameter space Θ of a parametric family $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$. In this section, we cover the function class $\mathcal{F}$ directly without parametrizing it. For this, though we need to measure the difference between two different functions f and g. However, in the previous section, we also noticed that in terms of generalization we only care about the function behavior on data. Therefore, if two functions behave the same over data and differently on other points, we treat these two functions as the same. The following difference metric makes this idea concrete.

Definition 37 (Difference metric)**.** *Given a dataset $\{z_1, ..., z_n\}$, we use the following to measure the difference between two functions.*

$$d(f, g) = \left(\frac{1}{n}\sum_{i=1}^{n}(f(z_i) - g(z_i))^2\right)^{1/2}$$

New notation: The following notation will simplify the statements and will be used throughout this section. Since we only care about function behavior on data, we can encode a function $f \in \mathcal{F}$ as a n-dimensional vector, i.e.,

$$\mathbf{f} = \tfrac{1}{\sqrt{n}}[f(z_1), f(z_2), ..., f(z_n)]^\top \in \mathbb{R}^n.$$

Using this new notation, we can simplify the following quantities as

$$\|\mathbf{f}\|^2 = \frac{1}{n}\sum_{i=1}^{n} f(z_i)^2 \quad \text{and} \quad d(f, g) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[f(z_i) - g(z_i)]^2} = \|\mathbf{f} - \mathbf{g}\|,$$

where the norms are understood to be Euclidean. We can also restate the Massart's Finite Lemma in this notation in a very compact form

$$\hat{\mathfrak{R}}_n(\mathcal{F}) \leq (\sup_{f\in\mathcal{F}} \|f\|)\sqrt{\frac{2\log(|\mathcal{F}|)}{n}}.$$

Reader should convince themselves that the above inequality is equivalent to the statement in Massart's Finite Lemma. We recall some of the key definitions of Section 5.

Definition 38. *We recall the following notions related to covering.*

- *ϵ-cover of $\mathcal{F}$ with respect to distance metric d is a set $\mathcal{N}_\epsilon = \{g_1, g_2, ...\}$ satisfying $\forall f \in \mathcal{F}$, $\exists g \in \mathcal{N}_\epsilon$ such that $d(f, g) \leq \epsilon$.*
- *Covering number of $\mathcal{F}$ is given by $N(\epsilon, \mathcal{F}, d) = \min\{|\mathcal{N}_\epsilon| : \mathcal{N}_\epsilon$ is a ϵ-cover of $\mathcal{F}\}$.*
- *Metric entropy of $\mathcal{F}$ is given by $\log N(\epsilon, \mathcal{F}, d)$.*

In general, one only needs an upper bound on the covering number. Therefore, our strategy will be to first construct a reasonable ϵ-cover of $\mathcal{F}$, then find an upper bound on its size which in turn upper bounds the covering number of $\mathcal{F}$. The following two examples demonstrate how to do this.

Example. [All functions $\mathbb{R} \to [0,1]$] Consider the function class $\mathcal{F} = \{f : \mathbb{R} \to [0,1]\}$. In order to cover this function class, we consider the 2d-grid defined by the points on the x-axis $z_{(1)}, z_{(2)}, ..., z_{(n)}$, the ordered data points, and the points on the y-axis $\{0, 2\epsilon, 4\epsilon, ...\}$. For each function $f \in \mathcal{F}$ and for each data point $z_{(i)}$ on the x-axis, we find the closest point on the grid and define a function g that passes on these points. It is easy to show that $d(f,g) \leq \epsilon$. Therefore if we include such functions g that pass on the points on this grid, we can obtain an ϵ-cover of $\mathcal{F}$. This suggests that we only need at most as many points as the number of points on this grid which can be upper bounded by

$$N(\epsilon, \mathcal{F}, d) \leq |\mathcal{N}_\epsilon| \leq (1 + 1/(2\epsilon))^n \leq (1/\epsilon)^n$$

for small ϵ. Notice that this number is exponential in n.

Example. [Non-decreasing function $\mathbb{R} \to [0,1]$] This time, consider the function class $\mathcal{F} = \{f : \mathbb{R} \to [0,1]$, and f non-decreasing$\}$. Using the same grid as before, we only need to count the number of non-decreasing functions that can be defined on this grid. This number can be upper bounded with $n^{1/\epsilon}$ which in turn implies

$$N(\epsilon, \mathcal{F}, d) \leq n^{1/\epsilon}.$$

We note that this bound is polynomial in n.

9.2 Simple discretization

In this section, we use an argument similar to Section 5 to obtain an upper bound on Rademacher complexity.

Theorem 39 (Discretization)**.** *For a function class $\mathcal{F} \subset \{f : \mathcal{Z} \to \mathbb{R}\}$, let $\kappa = \sup_{f \in \mathcal{F}} \|f\|$. Then,*

$$\forall \epsilon > 0, \quad \hat{\mathfrak{R}}_n(\mathcal{F}) \leq \kappa \sqrt{\frac{2 \log N(\epsilon, \mathcal{F}, d)}{n}} + \epsilon.$$

Remark. Before moving to proof, we make the following remarks.

1. Notice that with increasing ϵ, the first term in the right hand side of above bound decreases whereas the second term increases. This shows that there is a trade-off involving the parameter ϵ, the bound can be optimized over this parameter.

2. The above bound looks quite familiar. The first term above simply follows from the Massart's Finite Lemma whereas the second term is the discretization error.

Proof. Let $\boldsymbol{\sigma} = \frac{1}{\sqrt{n}}[\sigma_1, ..., \sigma_n]^\top$ be the vector of Rademacher random variables. We have $\|\boldsymbol{\sigma}\| = 1$, and also the empirical Rademacher complexity can be written as

$$\hat{\mathfrak{R}}_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \langle \boldsymbol{\sigma}, \mathbf{f} \rangle \big| z_{1:n}\right].$$

Let $\mathcal{N}_\epsilon$ be an ϵ-net over $\mathcal{F}$. Then $\forall f \in \mathcal{F}$, $\exists g \in \mathcal{N}_\epsilon$ such that $\|\mathbf{f} - \mathbf{g}\| \leq \epsilon$. Hence, we can write for any $f \in \mathcal{F}$

$$\begin{aligned}\langle \boldsymbol{\sigma}, \mathbf{f} \rangle =& \langle \boldsymbol{\sigma}, \mathbf{g} \rangle + \langle \boldsymbol{\sigma}, \mathbf{f} - \mathbf{g} \rangle \\ \leq& \langle \boldsymbol{\sigma}, \mathbf{g} \rangle + \|\boldsymbol{\sigma}\| \|\mathbf{f} - \mathbf{g}\| \quad \text{by Cauchy-Schwartz} \\ \leq& \max_{g \in \mathcal{N}_\epsilon} \langle \boldsymbol{\sigma}, \mathbf{g} \rangle + \epsilon.\end{aligned}$$

Now that the right hand side above doesn't depend on the choice of f, we can take supremum on the left hand side and obtain

$$\sup_{f \in \mathcal{F}} \langle \boldsymbol{\sigma}, \mathbf{f} \rangle \leq \max_{g \in \mathcal{N}_\epsilon} \langle \boldsymbol{\sigma}, \mathbf{g} \rangle + \epsilon.$$

Hence, we can write

$$\begin{aligned}\hat{\mathfrak{R}}_n(\mathcal{F}) \leq& \mathbb{E}\left[\max_{g \in \mathcal{N}_\epsilon} \langle \boldsymbol{\sigma}, \mathbf{g} \rangle\right] + \epsilon, \\ =& \mathfrak{R}_n(\mathcal{N}_\epsilon) + \epsilon, \\ \leq& (\sup_{g \in \mathcal{N}_\epsilon} \|g\|) \sqrt{\frac{2 \log |\mathcal{N}_\epsilon|}{n}} + \epsilon,\end{aligned}$$

which holds for all ϵ-covers of $\mathcal{F}$. Hence we can use the best cover and conclude the proof.

□

Using this theorem on the previous examples that we calculated an upper bound on the covering number, we can obtain an explicit rate.

Example.

1. All functions $\mathbb{R} \to [0, 1]$: We had the bound $N(\epsilon, \mathcal{F}, d) \leq (1/\epsilon)^n$. By the above theorem, we obtain

   $$\hat{\mathfrak{R}}_n(\mathcal{F}) \leq \sqrt{\frac{n \log(1/\epsilon)}{n}} + \epsilon = \mathcal{O}(1).$$

   We don't get generalization in this case.

2. Non-decreasing functions $\mathbb{R} \to [0, 1]$: We had the bound $N(\epsilon, \mathcal{F}, d) \leq n^{1/\epsilon})$. By the above theorem, we obtain

   $$\hat{\mathfrak{R}}_n(\mathcal{F}) \leq \sqrt{\frac{\log(n)}{\epsilon n}} + \epsilon.$$

   Optimizing over ϵ yields the rate $\mathcal{O}\big((\log(n)/n)^{1/3}\big)$. We do get generalization in this case, but this rate is slow, and interestingly it is just an artifact of the proof technique and can be improved.

9.3 Chaining

Next, we will see a more powerful technique called "chaining" which will improve the above rate significantly. We first state the main result of this section.

Theorem 40 (Dudley's Theorem)**.** *Let $\mathcal{F}$ be a set of functions $f : \mathcal{Z} \to \mathbb{R}$. Then,*

$$\hat{\mathfrak{R}}_n(\mathcal{F}) \leq 12 \int_0^\infty \sqrt{\frac{\log N(\epsilon, \mathcal{F}, d)}{n}} d\epsilon.$$

Remark. Before proving this theorem, we make a few remarks.

1. When the function class is composed of functions with finite norm, i.e. $\sup_{f \in \mathcal{F}} \|f\| = \kappa < \infty$, then the upper boundary of the above integral is κ since beyond that point covering number $N(\epsilon, \mathcal{F}, d) = 1$.

2. We notice that the discretization error in the result of Theorem 39 is gone!

3. For the above example on non-decreasing functions, since $\sup_{f \in \mathcal{F}} \|f\| = 1$, using the Dudley's theorem, we obtain

$$\hat{\mathfrak{R}}_n(\mathcal{F}) \leq 12 \int_0^1 \sqrt{\frac{\log(n)}{\epsilon n}} d\epsilon = \mathcal{O}\left(\sqrt{\frac{\log(n)}{n}}\right).$$

   This improves the previous rate of $\mathcal{O}\big((\log(n)/n)^{1/3}\big)$ significantly.

Proof. [by chaining]

Figure 3: Chaining idea (to be added)

Let's start by the most crude ϵ-cover for our function class, i.e. set $\epsilon_0 = \sup_{f \in \mathcal{F}} \|\mathbf{f}\|$ and note that we can set $\mathcal{N}_{\epsilon_0} = \{g_0\}$ for $g_0 = 0$ which implies $N(\epsilon_0, \mathcal{F}, d) = 1$. Next, define the sequence of epsilon covers $\mathcal{N}_{\epsilon_j}$ by setting $\epsilon_j = 2^{-j}\epsilon_0$. By definition, $\forall f \in \mathcal{F}$ we can find $g_j \in \mathcal{N}_{\epsilon_j}$ that depends on the choice of f such that $\|\mathbf{f} - \mathbf{g}_j\| \leq \epsilon$.

For any $m \in \mathbb{N}$, we can write the telescopic sum

$$\mathbf{f} = \mathbf{f} - \mathbf{g}_m + \sum_{j=1}^{m} \mathbf{g}_j - \mathbf{g}_{j-1} \tag{9.1}$$

since we have $\mathbf{g}_0 = 0$. By construction, the difference sequence $\mathbf{g}_j - \mathbf{g}_{j-1}$ forms a chain that gets smaller with j (since they also get closer to $\mathbf{f}$). That is, by triangle inequality, we have

$$\|\mathbf{g}_j - \mathbf{g}_{j-1}\| \leq \|\mathbf{g}_j - \mathbf{f}\| + \|\mathbf{f} - \mathbf{g}_{j-1}\| \leq \epsilon_j + \epsilon_{j-1} = 3\epsilon_j.$$

We have

$$\begin{aligned}
\hat{\mathfrak{R}}_n(\mathcal{F}) =& \mathbb{E}\Big[\sup_{f\in\mathcal{F}}\langle\boldsymbol{\sigma},\mathbf{f}\rangle\big|z_{1:n}\Big] = \mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big\{\underbrace{\langle\boldsymbol{\sigma},\mathbf{f}-\mathbf{g}_m\rangle}_{\leq\|\boldsymbol{\sigma}\|\|\mathbf{f}-\mathbf{g}_m\| \text{ by CS}} + \langle\boldsymbol{\sigma},\sum_{j=1}^m \mathbf{g}_j-\mathbf{g}_{j-1}\rangle\Big\}\big|z_{1:n}\Big] \quad \text{by (9.1)}\\
\leq& \epsilon_m + \mathbb{E}\Big[\sup_{f\in\mathcal{F}}\langle\boldsymbol{\sigma},\sum_{j=1}^m \mathbf{g}_j-\mathbf{g}_{j-1}\rangle\big|z_{1:n}\Big] \quad \text{by CS and } \mathcal{N}_{\epsilon_m}\text{'s net property}\\
\leq& \epsilon_m + \sum_{j=1}^m \mathbb{E}\Big[\sup_{f\in\mathcal{F}}\langle\boldsymbol{\sigma},\mathbf{g}_j-\mathbf{g}_{j-1}\rangle\big|z_{1:n}\Big] \quad \text{by } \sup\Sigma\leq\Sigma\sup\\
\leq& \epsilon_m + \sum_{j=1}^m \mathbb{E}\Big[\sup_{h\in\mathcal{H}_j}\langle\boldsymbol{\sigma},\mathbf{h}\rangle\big|z_{1:n}\Big] \quad \text{where } \mathcal{H}_j=\{g_j-g_{j-1}: g_j\in\mathcal{N}_{\epsilon_j}, g_{j-1}\in\mathcal{N}_{\epsilon_{j-1}}, \|\mathbf{g}_j-\mathbf{g}_{j-1}\|\leq 3\epsilon_j\}\\
\leq& \epsilon_m + \sum_{j=1}^m \Big(\sup_{h\in\mathcal{H}_j}\|\mathbf{h}\|\Big)\sqrt{\frac{2\log\left(|\mathcal{N}_{\epsilon_j}|^2\right)}{n}} \quad \text{by Massart's lemma and } |\mathcal{H}_j|\leq|\mathcal{N}_{\epsilon_j}||\mathcal{N}_{\epsilon_{j-1}}|\leq|\mathcal{N}_{\epsilon_j}|^2\\
\leq& \epsilon_m + 12\sum_{j=1}^m(\epsilon_j-\epsilon_{j+1})\sqrt{\frac{\log|\mathcal{N}_{\epsilon_j}|}{n}} \quad \text{since } \Big(\sup_{h\in\mathcal{H}_j}\|\mathbf{h}\|\Big)\leq 3\epsilon_j\leq 6(\epsilon_j-\epsilon_{j+1})\\
=& \epsilon_m + 12\sum_{j=1}^m\int_{\epsilon_{j+1}}^{\epsilon_j}\sqrt{\frac{\log|\mathcal{N}_{\epsilon_j}|}{n}}dt\\
\leq& \epsilon_m + 12\sum_{j=1}^m\int_{\epsilon_{j+1}}^{\epsilon_j}\sqrt{\frac{\log|\mathcal{N}_t|}{n}}dt \quad \text{since } t\in[\epsilon_{j+1},\epsilon_j]\\
\leq& \epsilon_m + 12\int_{\epsilon_m}^{\epsilon_0}\sqrt{\frac{\log|\mathcal{N}_t|}{n}}dt.
\end{aligned}$$

The result follows by letting $m\to\infty$ and noticing that the above bound holds for every ϵ_j-cover. □