

# Machine Learning Theory (CS 6783)

## Lecture 5 : Symmetrization and Infinite classes

### 1 Recap

Last class we showed that

$$\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \leq \sup_D \mathbb{E}_S \left[ \sup_{f \in \mathcal{F}} \left\{ \mathbb{E} [\ell(f(x), y)] - \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) \right\} \right]$$

This was using the Empirical Risk Minimizer (ERM)

1. When  $|\mathcal{F}| < \infty$ , using the above we showed that

$$\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \leq \sqrt{\frac{\log |\mathcal{F}|}{n}}$$

2. For countably infinite class we showed MDL bound and the algorithm based on this bound.
3. However the learning rate was not uniform over  $\mathcal{F}$

Can we get rate uniform over  $\mathcal{F}$  for infinite classes  $\mathcal{F}$  ?

### 2 Infinite Hypothesis Class : first attempt

As a first attempt, one can think of approximating the function class to desired accuracy by a finite number of representative elements. We call this a point-wise cover.

**Definition 1.** We say that set  $\mathcal{F}_\delta = \{\tilde{f}_1, \dots, \tilde{f}_N\}$  is an  $\delta$  point-wise cover for function class  $\mathcal{F}$  if  $\forall f \in \mathcal{F}$  there exists  $i \in [N]$  s.t.

$$\sup_{x,y} |\ell(f(x), y) - \ell(\tilde{f}_i(x), y)| \leq \delta$$

Further define  $N(\delta)$  to be the smallest  $N$  such that there exists an  $\delta$  cover of  $\mathcal{F}$  of cardinality at most  $N$ .

**Claim 1.** For any function class  $\mathcal{F}$ , we have that

$$\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \leq \inf_{\delta > 0} \left\{ 4\delta + \sqrt{\frac{\log N(\delta)}{n}} \right\}$$

*Proof.* Let  $\mathcal{F}_\delta = \{\tilde{f}_1, \dots, \tilde{f}_{N(\delta)}\}$  be an  $\delta$  cover for the function class  $\mathcal{F}$ . Further for every  $f \in \mathcal{F}$ , let  $i(f)$  correspond to the index of the element in  $\mathcal{F}_\delta$  that is  $\delta$  close to that  $f$ . Now note that,

$$\begin{aligned} \mathbb{E}_S \left[ \sup_{f \in \mathcal{F}} \left\{ \mathbb{E} [\ell(f(x), y)] - \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) \right\} \right] \\ \leq \mathbb{E}_S \left[ \sup_{i \in [N_\delta]} \left\{ \mathbb{E} [\ell(\tilde{f}_i(x), y)] - \frac{1}{n} \sum_{t=1}^n \ell(\tilde{f}_i(x_t), y_t) \right\} \right] \\ + \mathbb{E}_S \left[ \sup_{f \in \mathcal{F}} \left| \mathbb{E} [\ell(f(x), y)] - \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) - \mathbb{E} [\ell(\tilde{f}_{i(f)}(x), y)] + \frac{1}{n} \sum_{t=1}^n \ell(\tilde{f}_{i(f)}(x_t), y_t) \right| \right] \\ \leq \sqrt{\frac{\log N(\delta)}{n}} + 4\delta \end{aligned}$$

where the first term in the last inequality is by using the finite class bound and the second term is by using the definition of  $\delta$  cover as  $\tilde{f}_{i(f)}$  is  $\delta$  close to  $f$ . Since choice of  $\delta$  was arbitrary we can take the infimum over choices of  $\delta$  to conclude the proof.  $\square$

**Example : linear predictor, absolute loss, 1 dimension**

$$f(x) = f \cdot x, \quad \mathcal{F} = \mathcal{X} = [-1, 1], \quad \mathcal{Y} = [-1, 1], \quad \ell(y', y) = |y - y'|$$

$$N_\delta = \frac{2}{\delta}, \text{ Cover given by } f_1 = -1, f_2 = -1 + \delta, \dots, f_{N_\delta-1} = 1 - \delta, f_{N_\delta} = 1.$$

$$V_n^{\text{stat}}(\mathcal{F}) \leq \sqrt{\frac{\log n}{n}}$$

**Example : linear predictor/loss,  $d$  dimensions**

$$f(x) = \mathbf{f}^\top \mathbf{x}, \quad \mathcal{F} = \mathcal{X} = \{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\|_2 \leq 1\}, \quad \mathcal{Y} = [-1, 1], \quad \ell(y', y) = y \cdot y'$$

$$N_\delta = \Theta\left(\frac{2}{\delta}\right)^d$$

$$V_n^{\text{stat}}(\mathcal{F}) \leq \sqrt{\frac{d \log n}{n}}$$

**Example : thresholds**

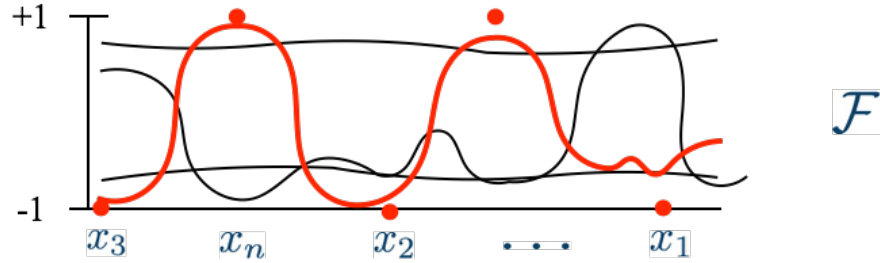
$$f(x) = \text{sign}(f - x), \quad \mathcal{F} = \mathcal{X} = [-1, 1], \quad \mathcal{Y} = \{-1, 1\}, \quad \ell(y', y) = \mathbf{1}_{\{y \neq y'\}}, \quad N_\delta = \infty \text{ for any } \delta < 1.$$

### 3 Symmetrization and Rademacher Complexity

$$\begin{aligned}
\mathbb{E}_S [L_D(\hat{y}_{\text{erm}})] - \inf_{f \in \mathcal{F}} L_D(f) &\leq \mathbb{E}_S \left[ \sup_{f \in \mathcal{F}} \left\{ \mathbb{E} [\ell(f(x), y)] - \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) \right\} \right] \\
&\leq \mathbb{E}_{S, S'} \left[ \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{t=1}^n \ell(f(x'_t), y'_t) - \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) \right\} \right] \\
&= \mathbb{E}_{S, S'} \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{t=1}^n \epsilon_t (\ell(f(x'_t), y'_t) - \ell(f(x_t), y_t)) \right\} \right] \\
&\leq 2 \mathbb{E}_S \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{t=1}^n \epsilon_t \ell(f(x_t), y_t) \right\} \right] \\
&=: \mathcal{R}_n(\mathcal{F})
\end{aligned}$$

Where in the above each  $\epsilon_t$  is a Rademacher random variable that is +1 with probability 1/2 and -1 with probability 1/2. The above is called Rademacher complexity of the loss class  $\ell \circ \mathcal{F}$ . In general Rademacher complexity of a function class measures how well the function class correlates with random signs. The more it can correlate with random signs the more complex the class is.

**Example :**  $\mathcal{X} = [0, 1]$ ,  $\mathcal{Y} = [-1, 1]$



### 4 Infinite $\mathcal{F}$ : Binary Classes and Growth Function

First let us simplify the Rademacher complexity for binary classification problem. Note that for binary classification problem where  $\mathcal{Y} \in \{\pm 1\}$ , the loss can be rewritten as

$\ell(y', y) = \mathbf{1}_{\{y \neq y'\}} = \frac{1 - y \cdot y'}{2}$ . Hence

$$\begin{aligned}
2 \mathbb{E}_S \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{t=1}^n \epsilon_t \ell(f(x_t), y_t) \right\} \right] &= 2 \mathbb{E}_S \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{t=1}^n \epsilon_t \frac{1 - f(x_t) \cdot y_t}{2} \right\} \right] \\
&= \mathbb{E}_S \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n \epsilon_t y_t f(x_t) \right]
\end{aligned}$$

Now consider the inner term in the expectation above, ie.  $\mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n \epsilon_t y_t f(x_t) \right]$ . Note that given any fixed choice of  $y_1, \dots, y_n \in \{\pm 1\}$ ,  $\epsilon_1 y_1, \dots, \epsilon_n y_n$  are also Rademacher random variables. Hence for the binary classification problem,

$$2\mathbb{E}_S \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{t=1}^n \epsilon_t \ell(f(x_t), y_t) \right\} \right] = \mathbb{E}_S \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n \epsilon_t f(x_t) \right]$$

In the above statement we moved from Rademacher complexity of loss class  $\ell \circ \mathcal{F}$  to the Rademacher complexity of the function class  $\mathcal{F}$  for binary classification task. This is a precursor to what we will refer to as contraction lemma which we will show later.

Why is symmetrization useful? Think what we gain for an infinite class ...