

# Introduction to Statistical Learning Theory

## Lecture 6



So far we have studied learnability via uniform convergence in binary classification and regression. We will now show a more general way to bound uniform convergence - Rademacher complexity.

First a small notation change - define  $z = (x, y)$  and  $l(h, z) = l(h(x), y)$ . This allows us to work in a more general setting with the same notation.

Another notation for simplicity: Define  $\mathcal{F} = l \circ \mathcal{H}$ , so for  $f \in \mathcal{F}$ -

$$L_{\mathcal{D}}(f) = \mathbb{E}_{\mathcal{D}}[f(z)] \text{ and } L_S(f) = \frac{1}{m} \sum_{i=1}^m f(z_i).$$

We are interested in bounding  $\sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_S(h))$ . As we have seen before, a good proxy for  $L_{\mathcal{D}}(h)$  is  $L_{\tilde{S}}(h)$  the loss on some second test sample. As we only have  $S$  we can split it into two equal size disjoint sets,  $S_1$  and  $S_2$ .



## Definition

$$\sup_{h \in \mathcal{H}} (L_{S_2}(h) - L_{S_1}(h)) = \frac{2}{m} \sup_{f \in \mathcal{F}} \left( \sum_{z_i \in S_2} f(z_i) - \sum_{z_j \in S_1} f(z_j) \right) =$$

$$\frac{2}{m} \sup_{f \in \mathcal{F}} \left( \sum_{i=1}^m \sigma_i f(z_i) \right). \quad \text{where } \sigma_i \in \{\pm 1\} \text{ indicates if } z_i \text{ is in } S_1 \text{ or } S_2. \text{ If}$$

we randomize  $\sigma_i$  we get the Rademacher complexity.

## Definition 1.1 (General Rademacher Complexity)

For  $A \subset \mathbb{R}^m$  define  $R(A) = \frac{1}{m} \mathbb{E}_{\sigma \sim \{\pm 1\}^m} \left[ \sup_{a \in A} \sum_{i=1}^m \sigma_i a_i \right]$

## Definition 1.2 (Empirical Rademacher Complexity)

Define  $\mathcal{F} \circ S = \{(f(z_1), \dots, f(z_m)) : f \in \mathcal{F}\} \subset \mathbb{R}^m$  the empirical Rademacher complexity is defined as

$$R(\mathcal{F} \circ S) = \frac{1}{m} \mathbb{E}_{\sigma \sim \{\pm 1\}^m} \left[ \sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i f(z_i) \right]$$





### Definition 1.3 (Rademacher Complexity )

The Rademacher complexity of  $\mathcal{F}$  is the expected empirical Rademacher complexity,  $\mathcal{R}_{\mathcal{D}}(\mathcal{F}, m) = \mathbb{E}_{S \sim \mathcal{D}^m} [R(\mathcal{F} \circ S)]$

The following lemma gives a nice intuition of the Rademacher complexity when considering binary classification

### Lemma 1.4

*Let  $\mathcal{H} : \mathcal{X} \rightarrow \{\pm 1\}$ ,  $S_X = \{x_1, \dots, x_m\} \subset \mathcal{X}$ . Define  $\text{err}(\mathcal{H})$  as the expected sample error of the ERM algorithm on random labels, then  $\text{err}(\mathcal{H}) = 1/2 - R(\mathcal{H} \circ S_X)/2$ .*



## Proof.

Let  $\sigma$  be any labeling on  $S_x$ . Then

$$\begin{aligned}L_{S_X, \sigma}(h) &= \frac{1}{m} \sum_{i=1}^m \mathbb{1}\{h(x_i) \neq \sigma_i\} = \frac{1}{m} \sum_{i=1}^m \frac{1 - \sigma_i h(x_i)}{2} \\ &= \frac{1}{2} - \frac{1}{2m} \sum_{i=1}^m \sigma_i h(x_i)\end{aligned}$$

This means that

$$L_{S_X, \sigma}(ERM) = \min_{h \in \mathcal{H}} \frac{1}{2} \left( 1 - \frac{1}{m} \sum_{i=1}^m \sigma_i h(x_i) \right) = \frac{1}{2} - \frac{1}{2} \max_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \sigma_i h(x_i).$$

Take expectation with regard to  $\sigma_i \sim \{\pm 1\}^m$  and you get the Rademacher complexity. □



## Lemma 1.5

$$\mathbb{E}_{S \sim \mathcal{D}^m} [\sup_{f \in \mathcal{F}} (L_{\mathcal{D}}(f) - L_S(f))] \leq 2\mathcal{R}_{\mathcal{D}}(\mathcal{F}, m)$$

Proof: Let  $\tilde{S} \sim \mathcal{D}^m$  be another sample, then

$$L_{\mathcal{D}}(f) - L_S(f) = \mathbb{E}_{\tilde{S}}[L_{\tilde{S}}(f)] - L_S(f) = \mathbb{E}_{\tilde{S}}[L_{\tilde{S}}(f) - L_S(f)] \quad (1)$$

therefore:

$$\sup_{f \in \mathcal{F}} (L_{\mathcal{D}}(f) - L_S(f)) = \sup_{f \in \mathcal{F}} \mathbb{E}_{\tilde{S}}[L_{\tilde{S}}(f) - L_S(f)] \quad (2)$$

$$\leq \mathbb{E}_{\tilde{S}}[\sup_{f \in \mathcal{F}} (L_{\tilde{S}}(f) - L_S(f))] \quad (3)$$



Taking expectation with regard to  $S$  we get

$$\mathbb{E}_{S \sim \mathcal{D}^m} [\sup_{f \in \mathcal{F}} (L_{\mathcal{D}}(f) - L_S(f))] \leq \mathbb{E}_{S, \tilde{S}} [\sup_{f \in \mathcal{F}} (L_{\tilde{S}}(f) - L_S(f))] \quad (4)$$

$$= \frac{1}{m} \mathbb{E}_{S, \tilde{S}} \left[ \sup_{f \in \mathcal{F}} \sum_{i=1}^m (f(\tilde{z}_i) - f(z_i)) \right] \quad (5)$$

$$= \frac{1}{m} \mathbb{E}_{S, \tilde{S}} \left[ \sup_{f \in \mathcal{F}} \sum_{i \neq j} (f(\tilde{z}_i) - f(z_i)) + f(\tilde{z}_j) - f(z_j) \right] \quad (6)$$

$$= \frac{1}{m} \mathbb{E}_{S, \tilde{S}} \left[ \sup_{f \in \mathcal{F}} \sum_{i \neq j} (f(\tilde{z}_i) - f(z_i)) + f(z_j) - f(\tilde{z}_j) \right] \quad (7)$$

$$= \frac{1}{m} \mathbb{E}_{S, \tilde{S}, \sigma_j} \left[ \sup_{f \in \mathcal{F}} \sum_{i \neq j} (f(\tilde{z}_i) - f(z_i)) + \sigma_j (f(z_j) - f(\tilde{z}_j)) \right] \quad (8)$$



This is true since  $z_i$  and  $\tilde{z}_i$  are drawn from the same distribution. We can do this for all  $1 \leq j \leq m$  and get

$$\mathbb{E}_{S \sim \mathcal{D}^m} [\sup_{f \in \mathcal{F}} (L_{\mathcal{D}}(f) - L_S(f))] \leq \frac{1}{m} \mathbb{E}_{S, \tilde{S}, \sigma} \left[ \sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i (f(\tilde{z}_i) - f(z_i)) \right] \quad (9)$$

$$\leq \frac{1}{m} \mathbb{E}_{S, \tilde{S}, \sigma} \left[ \sup_{f \in \mathcal{F}} \left( \sum_{i=1}^m \sigma_i f(\tilde{z}_i) \right) + \sup_{f \in \mathcal{F}} \left( \sum_{i=1}^m \sigma_i - f(z_i) \right) \right] \quad (10)$$

$$= 2\mathcal{R}_{\mathcal{D}}(\mathcal{F}, m) \quad \square \quad (11)$$



In order to turn our expectation bound to a high-probability bound, we need a concentration of measure theorem. We will use McDiarmid's inequality.

### Theorem 1.6 (McDiarmid's Inequality)

Let  $V$  be some set and  $f : V^m \rightarrow \mathbb{R}$  be a function such that for some  $c > 0$  and all  $x_1, \dots, x_m, x'_i \in V$  we have

$$|f(x_1, \dots, x_m) - f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_m)| \leq c \quad (12)$$

If  $X_1, \dots, X_m$  are independent r.v. taking values in  $V$ , then with probability greater or equal to  $1 - \delta$  we have

$$|f(X_1, \dots, X_m) - \mathbb{E}[f(X_1, \dots, X_m)]| \leq c \sqrt{\ln \left( \frac{2}{\delta} \right) \frac{m}{2}} \quad (13)$$



We can now state and prove the main theorem -

### Theorem 1.7

If for all  $z$  and  $h \in \mathcal{H}$  we have  $|l(h, z)| \leq c$ . Then with probability at least  $1 - \delta$ , for all  $h \in \mathcal{H}$ :

- 1  $|L_{\mathcal{D}}(h) - L_S(h)| \leq 2\mathcal{R}_{\mathcal{D}}(\mathcal{F}, m) + c\sqrt{\frac{2\ln(2/\delta)}{m}}$
- 2  $|L_{\mathcal{D}}(h) - L_S(h)| \leq 2R(\mathcal{F} \circ S) + 3c\sqrt{\frac{2\ln(4/\delta)}{m}}$
- 3  $L_{\mathcal{D}}(ERM) - L_{\mathcal{D}}(h^*) \leq 2R(\mathcal{F} \circ S) + 5c\sqrt{\frac{2\ln(8/\delta)}{m}}$

Notice that the last two inequalities only use the empirical sample, and can (up to computational complexity issues) be calculated for a given instance.



Proof:

We have  $\sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_S(h))$  satisfies the bounded differences condition eq. 12 with constant  $2c/m$ . Using the expectation bound of lemma 1.5 and the McDiarmid's inequality we have with probability  $\geq 1 - \delta$

$$\sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_S(h)) \leq \mathbb{E}_S[\sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_S(h))] + c\sqrt{\frac{2 \ln(2/\delta)}{m}} \quad (14)$$

$$\leq 2\mathcal{R}_{\mathcal{D}}(\mathcal{F}, m) + c\sqrt{\frac{2 \ln(2/\delta)}{m}}. \quad (15)$$

To prove the second inequality we note that  $\mathcal{R}_{\mathcal{D}}(\mathcal{F}, m)$  satisfies the bounded difference condition with the same constant, so with probability  $\geq 1 - \delta/2$ , we have  $\mathcal{R}_{\mathcal{D}}(\mathcal{F}, m) \leq R(\mathcal{F} \circ S) + c\sqrt{\frac{2 \ln(4/\delta)}{m}}$ . This and the union bound finish the proof of part 2.

The last part uses the 2nd inequality, the Hoeffding inequality to bound  $L_S(h^*) - L_{\mathcal{D}}(h^*)$  and the union bound. It is left as an exercise.



We will prove some useful theorem for bounding the Rademacher complexity.

### Lemma 2.1

*For any  $A \in \mathbb{R}^m$ , scalar  $c > 0$  and  $v \in \mathbb{R}^m$  we have  $R(cA + v) = cR(A)$ .*

Proof.

$$\begin{aligned} R(cA + v) &= \frac{1}{m} \mathbb{E}_\sigma \left[ \sup_{a \in A} \sum_{i=1}^m \sigma_i (ca_i + v_i) \right] = \\ &= \frac{1}{m} \mathbb{E}_\sigma \left[ c \sup_{a \in A} \left( \sum_{i=1}^m \sigma_i a_i \right) + \sum_{i=1}^m \sigma_i v_i \right] = cR(A) + \frac{1}{m} \sum_{i=1}^m v_i \mathbb{E}_\sigma [\sigma_i] = cR(A) \end{aligned}$$





## Lemma 2.2

For any  $A \in \mathbb{R}^m$ ,  $R(\text{conv}(A)) = R(A)$ , when  $\text{conv}(A)$  is the convex hull of  $A$ .

### Proof.

Define  $\Delta^n = \{\lambda \in \mathbb{R}^n : \forall i : \lambda_i \geq 0, \sum_{i=1}^n \lambda_i = 1\}$ . The convex hull is defined as  $\text{conv}(A) = \{\sum_{i=1}^n \lambda_i a^{(i)} : \forall i a^{(i)} \in A, \lambda \in \Delta^n\}$ . The key observation is that for every vector  $a$  we have  $\sup_{\lambda \in \Delta^n} \sum \lambda_i x_i = \max_i x_i$ .

$$\begin{aligned}
 mR(\text{conv}(A)) &= \mathbb{E}_\sigma \left[ \sup_{\lambda \in \Delta^n} \sup_{a^{(1)}, \dots, a^{(n)} \in A} \sum_{i=1}^m \sigma_i \sum_{j=1}^n \lambda_j a_i^{(j)} \right] \\
 &= \mathbb{E}_\sigma \left[ \sup_{\lambda \in \Delta^n} \sum_{j=1}^n \lambda_j \sup_{a^{(j)}} \sum_{i=1}^m \sigma_i a_i^{(j)} \right] = \mathbb{E}_\sigma \left[ \sup_{a \in A} \sum_{i=1}^m \sigma_i a_i \right] = R(A)
 \end{aligned}$$



## Lemma 2.3 (Massart Lemma)

If  $A = \{a_1, \dots, a_M\} \in \mathbb{R}^m$  is a finite set and  $\bar{a} = \frac{1}{N} \sum_{i=1}^N a_i$ , then

$$R(A) \leq \max_{a \in A} \|a - \bar{a}\| \frac{\sqrt{2 \ln(N)}}{m} \quad (16)$$

Immediate corollary - if  $l$  is the zero one loss and  $\mathcal{H}$  has VC dimension  $d$  then  $R(l \circ \mathcal{H} \circ S) \leq \sqrt{\frac{2d \ln(em/d)}{m}}$ .

Proof: Using lemma 2.1, we can assume  $\bar{a} = 0$  and recall that  $R(\lambda A) = \lambda R(A)$  for  $\lambda > 0$ .

$$\begin{aligned}
mR(\lambda A) &= \mathbb{E}_\sigma \left[ \max_{b \in \lambda A} \langle \sigma, b \rangle \right] = \mathbb{E}_\sigma \left[ \ln \left( \max_{b \in \lambda A} e^{\langle \sigma, b \rangle} \right) \right] \\
&\leq \mathbb{E}_\sigma \left[ \ln \left( \sum_{b \in \lambda A} e^{\langle \sigma, b \rangle} \right) \right] \stackrel{1}{\leq} \ln \left( \mathbb{E}_\sigma \left[ \sum_{b \in \lambda A} e^{\langle \sigma, b \rangle} \right] \right) \\
&\stackrel{2}{=} \ln \left( \sum_{b \in \lambda A} \prod \mathbb{E}_{\sigma_i} \left[ e^{\sigma_i \cdot b_i} \right] \right) \stackrel{3}{\leq} \ln \left( \sum_{b \in \lambda A} \prod e^{b_i^2/2} \right)
\end{aligned}$$

Where (1) is the Jensen inequality, (2) is from independence, and (3) is from a technical inequality  $\frac{e^a + e^{-a}}{2} \leq e^{a^2/2}$  we will prove shortly. We now have

$$\lambda mR(A) = mR(\lambda A) \leq \ln \left( \sum_{b \in \lambda A} e^{\|b\|^2/2} \right) \leq \ln(|A|) + \max_{a \in A} \lambda^2 \|a\|^2/2$$



We have  $R(A) \leq \frac{\ln(|A|) + \lambda^2 \max_{a \in A} (\|a\|^2/2)}{\lambda m}$ . Setting  $\lambda = \sqrt{2 \ln(|A|) / \max_{a \in A} \|a\|^2}$  concludes the proof.

We still need to prove  $\frac{e^a + e^{-a}}{2} \leq e^{a^2/2}$ . Using the Taylor series  $e^a = \sum_{i=0}^{\infty} \frac{a^i}{i!}$  so  $\frac{e^a + e^{-a}}{2} = \sum_{i=0}^{\infty} \frac{a^{2i}}{(2i)!}$ . On the other hand  $e^{a^2/2} = \sum_{i=0}^{\infty} \frac{a^{2i}}{2^i i!}$ . Observing that  $(2i)! \geq 2^i i!$  finishes the proof. □



## Lemma 2.4 (Contraction Lemma)

For all  $i \in [m]$ , let  $\phi_i : \mathbb{R} \rightarrow \mathbb{R}$  be a  $\rho$ -Lipschitz function. For all  $a \in \mathbb{R}^m$ , define  $\phi(a) = (\phi_1(a_1), \dots, \phi_m(a_m))$ . Then  $R(\phi \circ A) \leq \rho R(A)$ .

Using lemma 2.1 we can assume  $\rho = 1$ . Moreover if we define  $A_i = \{(a_1, \dots, a_{i-1}, \phi(a_i), a_{i+1}, \dots, a_m) : a \in A\}$ , it is enough to show  $R(A_1) \leq R(A)$ .

$$mR(A_1) = \mathbb{E}_\sigma \left[ \sup_{a \in A_1} \sum_{i=1}^m \sigma_i a_i \right] = \mathbb{E}_\sigma \left[ \sup_{a \in A} \sigma_1 \phi(a_1) + \sum_{i=2}^m \sigma_i a_i \right] \quad (17)$$

$$= \frac{1}{2} \mathbb{E}_{\sigma_2, \dots, \sigma_m} \left[ \sup_{a \in A} \left( \phi(a_1) + \sum_{i=2}^m \sigma_i a_i \right) + \sup_{a \in A} \left( -\phi(a_1) + \sum_{i=2}^m \sigma_i a_i \right) \right] \quad (18)$$

$$= \frac{1}{2} \mathbb{E}_{\sigma_2, \dots, \sigma_m} \left[ \sup_{a, a' \in A} \left( \phi(a_1) - \phi(a'_1) + \sum_{i=2}^m \sigma_i (a_i + a'_i) \right) \right] \quad (19)$$





$$\leq \frac{1}{2} \mathbb{E}_{\sigma_2, \dots, \sigma_m} \left[ \sup_{a, a' \in A} \left( |a_1 - a'_1| + \sum_{i=2}^m \sigma_i (a_i + a'_i) \right) \right] \quad (20)$$

$$= \frac{1}{2} \mathbb{E}_{\sigma_2, \dots, \sigma_m} \left[ \sup_{a, a' \in A} \left( a_1 - a'_1 + \sum_{i=2}^m \sigma_i (a_i + a'_i) \right) \right] \quad (21)$$

Where the last equality is from the fact we can switch  $a$  and  $a'$ . If we look at steps 17 – 19 using  $\phi = Id$  we see that the last line is equal to  $R(A)$  finishing the proof  $\square$ .