3.1 Introduction

In the last lecture we introduced the notion of Rademacher complexity and showed that it yields an upper bound on the expected value of the uniform (over the choice of actions/rules) deviation between the expected risk \( r \) and the empirical risk \( R \), namely,

\[
\mathbb{E} \sup_{a \in A} \{ r(a) - R(a) \} \leq 2 \mathbb{E} \text{Rad}(L \circ \{ Z_1, \ldots, Z_n \})
\]

where we recall the notation

\[
L := \{ z \in Z \rightarrow \ell(a, z) \in \mathbb{R} : a \in A \}.
\]

In this lecture we establish bounds for \( \text{Rad}(L \circ \{ z_1, \ldots, z_n \}) \) for any \( z_1, \ldots, z_n \in Z \) in the setting of regression.

In supervised learning, the observed examples correspond to pairs of points, i.e., \( Z_i = (X_i, Y_i) \in X \times Y \). The point \( X_i \) is called feature or covariate, and the point \( Y_i \) is its corresponding label. The set of admissible decisions is a subset of the set functions from \( X \) to \( Y \), i.e., \( A \subseteq B := \{ a : X \rightarrow Y \} \), and the loss function is of the form \( \ell(a, (x, y)) = \phi(a(x), y) \), for a function \( \phi : Y \times Y \rightarrow \mathbb{R}_+ \).

The regression setting is represented by the choice \( X = \mathbb{R}^d \) for a given dimension \( d \), \( Y = \mathbb{R} \). We have \( S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \) and \( s = \{(x_1, y_1), \ldots, (x_n, y_n)\} \) which represents a realization of the training sample. Let us recall the following notation:

\[
A \circ \{ x_1, \ldots, x_n \} := \{(a(x_1), \ldots, a(x_n)) \in Y^n : a \in A \}.
\]

The following proposition shows how to use the contraction property of Rademacher complexity, Lemma 2.10, to relate the Rademacher complexity of the set of interest (which involves the loss function \( \ell \)) to the Rademacher complexity of a set that only depends on \( A \). The main idea is that we want to be able to relate the quantity \( \mathbb{E} \sup_{a \in A} \{ r(a) - R(a) \} \) to something that depends on a notion of complexity for \( A \), for a general class of loss functions (loss functions that are Lipschitz).

**Proposition 3.1** Let the function \( \hat{y} \rightarrow \phi(\hat{y}, y) \) be \( \gamma \)-Lipschitz for any \( y \in Y \).

Then, for any \( (x_1, y_1), \ldots, (x_n, y_n) \in X \times Y \),

\[
\text{Rad}(L \circ \{(x_1, y_1), \ldots, (x_n, y_n)\}) \leq \gamma \text{Rad}(A \circ \{x_1, \ldots, x_n\})
\]

**Proof:** By the contraction property of Rademacher complexity, Lemma 2.10, we get

\[
\text{Rad}(L \circ s) = \mathbb{E} \sup_{w \in \mathbb{R}^d : \|w\|_2 \leq c} \frac{1}{n} \sum_{i=1}^n \Omega_i \phi(w^\top x_i, y_i) = \text{Rad}((\phi(\cdot, y_1), \ldots, \phi(\cdot, y_n)) \circ A \circ \{x_1, \ldots, x_n\}) \leq \gamma \text{Rad}(A \circ \{x_1, \ldots, x_n\}).
\]

Below we show how to control the quantity \( \text{Rad}(A \circ \{x_1, \ldots, x_n\}) \) for some function classes \( A \) of interest.
3.2 Linear predictors $\ell_2/\ell_2$ constraints

In the case of $\ell_2/\ell_2$ constraints, the Rademacher complexity of linear predictors does not depend explicitly on the dimension $d$ (the dependence on $d$ is implicit, via the term $\max_i \|x_i\|_2$).

**Proposition 3.2** Let $A_2 := \{x \in \mathbb{R}^d \rightarrow w^T x : w \in \mathbb{R}^d, \|w\|_2 \leq 1\}$. Then, for any $x_1, \ldots, x_n \in \mathbb{R}^d$,

$$\text{Rad}(A_2 \circ \{x_1, \ldots, x_n\}) \leq \frac{\max_i \|x_i\|_2}{\sqrt{n}}$$

**Proof:** We have

$$n \text{Rad}(A_2 \circ \{x_1, \ldots, x_n\}) = \mathbb{E} \sup_{w \in \mathbb{R}^d : \|w\|_2 \leq 1} \sum_{i=1}^n \Omega_i w^T x_i = \mathbb{E} \sup_{w \in \mathbb{R}^d : \|w\|_2 \leq 1} w^T \left(\sum_{i=1}^n \Omega_i x_i\right)$$

$$\leq \sup_{w \in \mathbb{R}^d : \|w\|_2 \leq 1} \|w\|_2 \mathbb{E} \left\|\sum_{i=1}^n \Omega_i x_i\right\|_2$$

by Cauchy-Schwarz’s ineq. $x^T y \leq \|x\|_2 \|y\|_2$

$$\leq \mathbb{E} \left(\sum_{i=1}^n \Omega_i x_i\right)_2^2 \leq \sqrt{n} \mathbb{E} \left(\sum_{i=1}^n \Omega_i x_i\right)_2^2$$

by Jensen’s, as $x \rightarrow \sqrt{x}$ is concave

$$= \sqrt{\mathbb{E} \sum_{j=1}^d \left(\sum_{i=1}^n \Omega_i x_{i,j}\right)^2}$$

$$= \sqrt{\mathbb{E} \sum_{j=1}^d \sum_{i=1}^n (\Omega_i x_{i,j})^2}$$

as the $\Omega_i$’s are independent and $\mathbb{E} \Omega_i = 0$

$$= \sqrt{\mathbb{E} \sum_{i=1}^n \|x_i\|_2^2} \leq \sqrt{n} \max_i \|x_i\|_2$$

as $\Omega_i^2 = 1$.

**Remark 3.3** Note that as the predictors that we are considering are linear, i.e., $x \in \mathbb{R}^d \rightarrow w^T x$, the constraint $\|w\|_2 \leq 1$ in the definition of $A_2$ in Proposition 3.2 is without loss of generality. In fact, if $\|w\|_2 \leq c$ for a given constant $c \geq 0$, then we can rescale $w^T x = (\frac{w}{\|w\|_2})^T (\|w\|_2 x)$ and we have the equivalence

$$\{x \in \mathbb{R}^d \rightarrow w^T x : w \in \mathbb{R}^d, \|w\|_2 \leq c\} = \{x \in \mathbb{R}^d \rightarrow w^T (cx) : w \in \mathbb{R}^d, \|w\|_2 \leq 1\}.$$

Proposition 3.2 still applies, with a constant $c$ on the right-hand side of the bound.

3.3 Linear predictors $\ell_1/\ell_\infty$ constraints ($\ell_1$ Boosting)

In the case of $\ell_1/\ell_\infty$ constraints, the Rademacher complexity of linear predictors only depends logarithmically on the dimension $d$.

**Proposition 3.4** Let $A_1 := \{x \in \mathbb{R}^d \rightarrow w^T x : w \in \mathbb{R}^d, \|w\|_1 \leq 1\}$. Then, for any $x_1, \ldots, x_n \in \mathbb{R}^d$,

$$\text{Rad}(A_1 \circ \{x_1, \ldots, x_n\}) \leq \frac{\max_i \|x_i\|_\infty}{\sqrt{n}} \sqrt{2 \log(2d)}$$
Proof: We have

\[ n \operatorname{Rad}(A_1 \circ \{x_1, \ldots, x_n\}) = \mathbb{E} \sup_{\omega \in \mathbb{R}^d : \|\omega\|_1 \leq 1} \sum_{i=1}^{n} \Omega_i \omega^\top x_i = \mathbb{E} \sup_{\omega \in \mathbb{R}^d : \|\omega\|_1 \leq 1} \omega^\top \left( \sum_{i=1}^{n} \Omega_i x_i \right) \]

\[ \leq \sup_{\omega \in \mathbb{R}^d : \|\omega\|_1 \leq 1} \left\| \omega \right\|_1 \mathbb{E} \left\| \sum_{i=1}^{n} \Omega_i x_i \right\|_\infty \text{ by Hölder’s inequality } x^\top y \leq \|x\|_1 \|y\|_\infty \]

\[ \leq \mathbb{E} \left\| \sum_{i=1}^{n} \Omega_i x_i \right\|_\infty \]

Let \( t_j := (x_{1,j}, \ldots, x_{n,j}) \in \mathbb{R}^n \) for any \( j \in 1 : d \), and let \( T = \{t_1, \ldots, t_d\} \). Then,

\[ \left\| \sum_{i=1}^{n} \Omega_i x_i \right\|_\infty = \max_{j \in 1 : d} \left\| \sum_{i=1}^{n} \Omega_i x_{1,j} \right\| = \max_{j \in 1 : d} \left\| \sum_{i=1}^{n} \Omega_i t_{j,i} \right\| = \max_{t \in T} \left\| \sum_{i=1}^{n} \Omega_i t_i \right\|, \]

whose expectation looks like a Rademacher complexity apart from the absolute value (and the normalization by \( 1/n \)). To remove the absolute value, note that for any \( \omega_1, \ldots, \omega_n \in \{-1, 1\}^n \) we have

\[ \max_{t \in T} \left\| \sum_{i=1}^{n} \omega_i t_i \right\| = \max_{t \in T \cup T_-} \left\| \sum_{i=1}^{n} \omega_i t_i \right\|, \]

where we have defined \( T_- = \{-t_1, \ldots, -t_d\} \), with \( -t_j = (-x_{1,j}, \ldots, -x_{n,j}) \). Hence, we have

\[ \operatorname{Rad}(A_1 \circ \{x_1, \ldots, x_n\}) \leq \operatorname{Rad}(T \cup T_-) , \]

and the proof follows by Massart’s lemma as

\[ \operatorname{Rad}(T \cup T_-) \leq \max_{t \in T \cup T_-} \left\| t \right\|_2 \sqrt{2 \log |T \cup T_-|} \leq \sqrt{n} \max_i \|x_i\|_\infty \sqrt{2 \log(2d)} \frac{n}{n}. \]

Remark 3.5 Note that as the predictors that we are considering are linear, i.e., \( x \in \mathbb{R}^d \rightarrow w^\top x \), the constraint \( \|w\|_1 \leq 1 \) in the definition of \( A_1 \) is without loss of generality. In fact, if \( \|w\|_1 \leq c \) for a given constant \( c \geq 0 \), then we can rescale \( w^\top x = (\frac{w}{\|w\|_1})^\top (\|w\|_1 x) \) and we have the equivalence

\[ \{x \in \mathbb{R}^d \rightarrow w^\top x : w \in \mathbb{R}^d, \|w\|_1 \leq c\} = \{x \in \mathbb{R}^d \rightarrow w^\top (cx) : w \in \mathbb{R}^d, \|w\|_1 \leq 1\} . \]

Proposition 3.4 still applies, with a constant \( c \) on the right-hand side of the bound.

3.4 Linear predictors simple\( \_\)x/\( \ell_\infty \) constraints (Boosting)

Proposition 3.6 Let \( \Delta_d := \{w \in \mathbb{R}^d : \|w\|_1 = 1, w_1, \ldots, w_d \geq 0\} \) and let \( A_\Delta := \{x \in \mathbb{R}^d \rightarrow w^\top x : w \in \Delta_d\} \). Then, for any \( x_1, \ldots, x_n \in \mathbb{R}^d \),

\[ \operatorname{Rad}(A_\Delta \circ \{x_1, \ldots, x_n\}) \leq \frac{\max_i \|x_i\|_\infty}{\sqrt{n}} \sqrt{2 \log d} \]
Proof: We have
\[ n \text{Rad}(A_\Delta \circ \{x_1, \ldots, x_n\}) = E \sup_{w \in \Delta_d} \sum_{i=1}^{n} \Omega_i w^\top x_i = E \sup_{w \in \Delta_d} w^\top \left( \sum_{i=1}^{n} \Omega_i x_i \right). \]

Note that for any vector \( v = (v_1, \ldots, v_d) \in \mathbb{R}^d \) we have
\[ \sup_{w \in \Delta_d} w^\top v = \max_{j \in 1:d} v_j. \]

Then,
\[ E \sup_{w \in \Delta_d} w^\top \left( \sum_{i=1}^{n} \Omega_i x_i \right) = E \max_{j \in 1:d} \sum_{i=1}^{n} \Omega_i x_{i,j} = n \text{Rad}(T), \]

with \( T = \{t_1 \ldots, t_d\} \) with \( t_j = (x_{1,j}, \ldots, x_{n,j}) \) for any \( j \in \{1, \ldots, d\} \). The proof follows by Massart’s lemma as
\[ \text{Rad}(T) \leq \max_{t \in T} \|t\|_2 \frac{\sqrt{2 \log |T|}}{n} \leq \sqrt{n} \max_{i} \|x_i\|_\infty \frac{\sqrt{2 \log d}}{n}. \]

3.5 Feed-forward neural networks

Let us define a feed-forward neural networks with activation functions applied element-wise to its units.

A layer \( l^{(k)} : \mathbb{R}^{d_{k-1}} \rightarrow \mathbb{R}^{d_k} \) consists of a coordinate-wise composition of an activation function \( \sigma^{(k)} : \mathbb{R} \rightarrow \mathbb{R} \) and an affine map, namely,
\[ l^{(k)}(x) := \sigma^{(k)}(W^{(k)} x + b^{(k)}), \]

for a given interaction matrix \( W^{(k)} \) and bias vector \( b^{(k)} \).

A neural network with depth \( p \) (and \( p-1 \) hidden layers) is given by the function \( f_{nn}^{(p)} : \mathbb{R}^d \rightarrow \mathbb{R} \) defined as
\[ f_{nn}^{(p)}(x) := l^{(p)} \circ l^{(p-1)} \circ \cdots \circ l^{(1)}(x) = l^{(p)}(\cdots(l^{(2)}(l^{(1)}(x)))) \cdots, \]

with \( d_0 = d, d_p = 1 \), \( \sigma^{(r)} = \sigma \) for a given function \( \sigma \) for all \( r < p \), and \( \sigma^{(p)}(x) = x \) (i.e., the last layer is simply an affine map).

The activation function \( \sigma \) is known to the design maker, while the interaction matrices and the bias vectors are treated as parameters to tune. For instance, a class of neural networks with depth \( p \) is given by
\[ A_{mn}^{(p)} := \{ x \in \mathbb{R}^d \rightarrow f_{nn}^{(p)}(x) : \|W^{(k)}\|_\infty \leq \omega, \|b^{(k)}\|_\infty \leq \beta \ \forall k \}, \]

where for a given matrix \( M \), the \( \ell_\infty \) norm is defined as \( \|M\|_\infty := \max_{i} \sum_{j} |M_{ij}| \).

The Rademacher complexity of a feed-forward neural network can be bounded recursively by considering each layer at a time. A bound that can be used for the recursion is given by the following proposition, that expresses the Rademacher complexities at the outputs of one layer in terms of the outputs at the previous layers.

**Proposition 3.7** Let \( \mathcal{L} \) be a class of functions from \( \mathbb{R}^d \) to \( \mathbb{R} \) that includes the zero function. Let \( \sigma : \mathbb{R} \rightarrow \mathbb{R} \) be \( \alpha \)-Lipschitz and define \( \mathcal{L}' := \{ x \in \mathbb{R}^d \rightarrow \sigma(\sum_{j=l_1}^{l_m} w_j l_j(x) + b) \in \mathbb{R} : \|b\|_1 \leq \beta, \|w\|_1 \leq \omega, l_1, \ldots, l_m \in \mathcal{L} \} \).

Then, for any \( x_1, \ldots, x_n \in \mathbb{R}^d \),
\[ \text{Rad}(\mathcal{L}' \circ \{x_1, \ldots, x_n\}) \leq \alpha \left( \frac{\beta}{\sqrt{n}} + 2\omega \text{Rad}(\mathcal{L} \circ \{x_1, \ldots, x_n\}) \right) \]

(3.2)
If \( \mathcal{L} = -\mathcal{L} \), then

\[
\text{Rad}(\mathcal{L}' \circ \{x_1, \ldots, x_n\}) \leq \alpha \left( \frac{\beta}{\sqrt{n}} + \omega \text{Rad}(\mathcal{L} \circ \{x_1, \ldots, x_n\}) \right)
\] (3.3)

**Proof:** We give a proof that makes use of many of the property of the Rademacher complexity described in the previous lecture. Let

\[
\mathcal{F} : = \{ x \in \mathbb{R}^d \to \sum_{i=1}^{m} w_i l_j(x) \in \mathbb{R} : \|w\|_1 \leq \omega, l_1, \ldots, l_m \in \mathcal{L} \},
\]

\[
\mathcal{G} : = \{ x \in \mathbb{R}^d \to b \in \mathbb{R} : |b| \leq \beta \}.
\]

By the contraction property and the summation property of Rademacher complexities, we have

\[
\text{Rad}(\mathcal{L}' \circ \{x_1, \ldots, x_n\}) \leq \alpha \left( \text{Rad}(\mathcal{F} \circ \{x_1, \ldots, x_n\}) + \text{Rad}(\mathcal{G} \circ \{x_1, \ldots, x_n\}) \right).
\]

On the one hand, as \( \mathcal{L} \) contains the zero function we have

\[
\mathcal{F} \circ \{x_1, \ldots, x_n\} = \omega \text{ conv}(\mathcal{L} - \mathcal{L}),
\]

where \( \mathcal{L} - \mathcal{L} = \{ l - l' : l \in \mathcal{L}, l' \in \mathcal{L}' \} \). In fact, first of all note that

\[
\text{Rad}(\mathcal{F} \circ \{x_1, \ldots, x_n\}) = \text{Rad}(\mathcal{F}' \circ \{x_1, \ldots, x_n\})
\]

where

\[
\mathcal{F}' := \{ x \in \mathbb{R}^d \to \sum_{i=1}^{m} w_i l_j(x) \in \mathbb{R} : \|w\|_1 = \omega, l_1, \ldots, l_m \in \mathcal{L} \}
\]

(this is because the maximum over \( \|w\|_1 \leq \omega \) is achieved for the values satisfying \( \|w\|_1 = \omega \)). Then, note that for any \( w \in \mathbb{R}^m \) such that \( \|w\|_1 = 1 \) we have

\[
\sum_{i} w_i l_i = \sum_{i: w_i \geq 0} w_i (l_i - 0) + \sum_{i: w_i < 0} |w_i|(0 - l_i),
\]

where 0 represents the zero function. The right-hand side is a convex combination of elements in \( \mathcal{L} - \mathcal{L}' \). Hence, by the convex hull property of Rademacher complexity we find

\[
\text{Rad}(\mathcal{F} \circ \{x_1, \ldots, x_n\}) = \omega \text{ Rad}(\text{ conv}(\mathcal{L} - \mathcal{L}) \circ \{x_1, \ldots, x_n\}) = \omega \text{ Rad}(\mathcal{L} \circ \{x_1, \ldots, x_n\})
\]

\[
= \omega \text{ Rad}(\mathcal{L} \circ \{x_1, \ldots, x_n\}) + \omega \text{ Rad}(\mathcal{L} \circ \{x_1, \ldots, x_n\}) = 2 \omega \text{ Rad}(\mathcal{L} \circ \{x_1, \ldots, x_n\}),
\]

where the factor 2 is not necessary if \( \mathcal{L} = -\mathcal{L} \). On the other hand,

\[
n \text{Rad}(\mathcal{G} \circ \{x_1, \ldots, x_n\}) = E \sup_{b : |b| \leq \beta} b \left| \sum_{i=1}^{n} \Omega_i \leq E \sup_{|b| \leq \beta} |b| \left| \sum_{i=1}^{n} \Omega_i \right| \leq \beta \sqrt{n},
\]

where the last inequality follows by Jensen's inequality, as \( E[\sum_{i=1}^{n} \Omega_i] \leq \sqrt{E[(\sum_{i=1}^{n} \Omega_i)^2]} = \sqrt{n} \) using the independence of the \( \Omega_i \)’s and that \( \Omega_i^2 = 1 \).

We are now ready to give a bound for the full neural network. We can use Proposition 3.7 to run the recursion, noticing that the last layer involves a linear function (which is 1-Lipschitz). The first layer requires a different treatment, and we can use Proposition 3.4.

Using Proposition 3.7 we can establish the following bound for the Rademacher complexity of a layered neural network.
Proposition 3.8 Let $\sigma$ be $\lambda$-Lipschitz. Let $A_{nn}^{(p)}$ be defined as in 3.1. Then, for any $x_1, \ldots, x_n \in \mathbb{R}^d$,

$$\text{Rad}(A_{nn}^{(p)} \circ \{x_1, \ldots, x_n\}) \leq \frac{1}{\sqrt{n}} \left( \beta + 2\omega \beta \lambda \sum_{k=0}^{p-3} (2\omega \lambda)^k + 2\omega (2\omega \lambda)^{p-2} \max_i \|x_i\|_\infty \sqrt{2 \log(2d)} \right)$$

If $\lambda = 1$ and $\sigma$ is anti-symmetric, namely, $\sigma(x) = -\sigma(-x)$, we have

$$\text{Rad}(A_{nn}^{(p)} \circ \{x_1, \ldots, x_n\}) \leq \frac{1}{\sqrt{n}} \left( \beta \sum_{k=0}^{p-2} \omega^k + \omega^{p-1} \max_i \|x_i\|_\infty \sqrt{2 \log(2d)} \right)$$

Proof: As the last layer of the neural network is linear, i.e., $\sigma^{(p)}(x) = x$, we can apply Proposition 3.7 with $\alpha = 1$ (as $\sigma^{(p)}$ is 1-Lipschitz) once and then apply (3.2) in Proposition 3.7 with $\alpha = \lambda$ for $p - 2$ times. We find

$$\text{Rad}(A_{nn}^{(p)} \circ \{x_1, \ldots, x_n\}) \leq \beta \sqrt{n} + 2\omega \left( \frac{\beta \lambda}{\sqrt{n}} \sum_{k=0}^{p-3} (2\omega \lambda)^k + (2\omega \lambda)^{p-2} \text{Rad}(A_1 \circ \{x_1, \ldots, x_n\}) \right).$$

The result of the first inequality follows by Proposition 3.4. The second inequality can be proved using the same strategy, using (3.3) instead of (3.2).