**Algorithmic Foundations of Learning** **Lecture 3**

# Rademacher Complexity. Examples

*Lecturer: Patrick Rebeschini* *Version: January 25, 2022*

## 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,

$$\boxed{\mathbf{E} \sup_{a \in \mathcal{A}} \{r(a) - R(a)\} \leq 2\, \mathbf{E}\, \mathtt{Rad}(\mathcal{L} \circ \{Z_1, \dots, Z_n\})}$$

where we recall the notation

$$\mathcal{L} := \{z \in \mathcal{Z} \to \ell(a, z) \in \mathbb{R} : a \in \mathcal{A}\}.$$

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

In supervised learning, the observed examples correspond to pairs of points, i.e., $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$. 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 $\mathcal{X}$ to $\mathcal{Y}$, i.e., $\mathcal{A} \subseteq \mathcal{B} := \{a : \mathcal{X} \to \mathcal{Y}\}$, and the loss function is of the form $\ell(a, (x, y)) = \phi(a(x), y)$, for a function $\phi : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$.

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

$$\mathcal{A} \circ \{x_1, \dots, x_n\} := \{(a(x_1), \dots, a(x_n)) \in \mathcal{Y}^n : a \in \mathcal{A}\}.$$

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 $\mathcal{A}$. The main idea is that we want to be able to relate the quantity $\mathbf{E} \sup_{a \in \mathcal{A}} \{r(a) - R(a)\}$ to something that depends on a notion of complexity for $\mathcal{A}$, for a general class of loss functions (loss functions that are Lipschitz).

**Proposition 3.1** *Let the function $\hat{y} \to \phi(\hat{y}, y)$ be $\gamma$-Lipschitz for any $y \in \mathcal{Y}$. Then, for any $(x_1, y_1), \dots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$,*

$$\boxed{\mathtt{Rad}(\mathcal{L} \circ \{(x_1, y_1), \dots, (x_n, y_n)\}) \leq \gamma\, \mathtt{Rad}(\mathcal{A} \circ \{x_1, \dots, x_n\})}$$

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

$$\begin{aligned} \mathtt{Rad}(\mathcal{L} \circ s) &= \mathbf{E} \sup_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \Omega_i \phi(a(x_i), y_i) = \mathtt{Rad}((\phi(\,\cdot\,, y_1), \dots, \phi(\,\cdot\,, y_n)) \circ \mathcal{A} \circ \{x_1, \dots, x_n\}) \\ &\leq \gamma\, \mathtt{Rad}(\mathcal{A} \circ \{x_1, \dots, x_n\}). \end{aligned}$$

■

Below we show how to control the quantity $\mathtt{Rad}(\mathcal{A} \circ \{x_1, \dots, x_n\})$ for some function classes $\mathcal{A}$ 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 grows as $\sqrt{d}$.

**Proposition 3.2** *Let $\mathcal{A}_2 := \{x \in \mathbb{R}^d \to w^\top x : w \in \mathbb{R}^d, \|w\|_2 \le c\}$ for a positive constant $c > 0$. Then, for any $x_1 \dots, x_n \in \mathbb{R}^d$,*

$$\boxed{\mathtt{Rad}(\mathcal{A}_2 \circ \{x_1, \dots, x_n\}) \le c\frac{\max_i \|x_i\|_2}{\sqrt{n}} \le c \max_i \|x_i\|_\infty \frac{\sqrt{d}}{\sqrt{n}}}$$

**Proof:** First, consider the case $c = 1$. We have

$$\begin{aligned}
n\,\mathtt{Rad}(\mathcal{A}_2 \circ \{x_1, \dots, x_n\}) &= \mathbf{E} \sup_{w \in \mathbb{R}^d : \|w\|_2 \le 1} \sum_{i=1}^n \Omega_i w^\top x_i = \mathbf{E} \sup_{w \in \mathbb{R}^d : \|w\|_2 \le 1} w^\top \Big( \sum_{i=1}^n \Omega_i x_i \Big) \\
&\le \sup_{w \in \mathbb{R}^d : \|w\|_2 \le 1} \|w\|_2\, \mathbf{E} \Big\| \sum_{i=1}^n \Omega_i x_i \Big\|_2 \quad \text{by Cauchy-Schwarz's ineq. } x^\top y \le \|x\|_2 \|y\|_2 \\
&\le \mathbf{E} \sqrt{\Big\| \sum_{i=1}^n \Omega_i x_i \Big\|_2^2} \le \sqrt{\mathbf{E} \Big\| \sum_{i=1}^n \Omega_i x_i \Big\|_2^2} \quad \text{by Jensen's, as } x \to \sqrt{x} \text{ is concave} \\
&= \sqrt{\mathbf{E} \sum_{j=1}^d \Big( \sum_{i=1}^n \Omega_i x_{i,j} \Big)^2} \\
&= \sqrt{\mathbf{E} \sum_{j=1}^d \sum_{i=1}^n (\Omega_i x_{i,j})^2} \quad \text{as the } \Omega_i\text{'s are independent and } \mathbf{E}\Omega_i = 0 \\
&= \sqrt{\mathbf{E} \sum_{i=1}^n \|x_i\|_2^2} \le \sqrt{n} \max_i \|x_i\|_2 \quad \text{as } \Omega_i^2 = 1.
\end{aligned}$$

If $c \neq 1$, then we can rescale $w^\top x = (\frac{w}{\|w\|_2})^\top (\|w\|_2 x)$ and the following equivalence concludes the proof:

$$\{x \in \mathbb{R}^d \to w^\top x : w \in \mathbb{R}^d, \|w\|_2 \le c\} = \{x \in \mathbb{R}^d \to w^\top (cx) : w \in \mathbb{R}^d, \|w\|_2 \le 1\}.$$

■

## 3.3 Linear predictors $\ell_1/\ell_\infty$ constraints

In the case of $\ell_1/\ell_\infty$ constraints, the Rademacher complexity of linear predictors only grows as $\sqrt{\log d}$.

**Proposition 3.3** *Let $\mathcal{A}_1 := \{x \in \mathbb{R}^d \to w^\top x : w \in \mathbb{R}^d, \|w\|_1 \le c\}$ for a positive constant $c > 0$. Then, for any $x_1 \dots, x_n \in \mathbb{R}^d$,*

$$\boxed{\mathtt{Rad}(\mathcal{A}_1 \circ \{x_1, \dots, x_n\}) \le c \max_i \|x_i\|_\infty \frac{\sqrt{2\log(2d)}}{\sqrt{n}}}$$

**Proof:** First, consider the case $c = 1$. We have

$$\begin{aligned} n\,\mathtt{Rad}(\mathcal{A}_1 \circ \{x_1, \dots, x_n\}) &= \mathbf{E} \sup_{w \in \mathbb{R}^d : \|w\|_1 \le 1} \sum_{i=1}^n \Omega_i w^\top x_i = \mathbf{E} \sup_{w \in \mathbb{R}^d : \|w\|_1 \le 1} w^\top \Big( \sum_{i=1}^n \Omega_i x_i \Big) \\ &\le \sup_{w \in \mathbb{R}^d : \|w\|_1 \le 1} \|w\|_1 \, \mathbf{E} \Big\| \sum_{i=1}^n \Omega_i x_i \Big\|_\infty \quad \text{by Hölder's inequality } x^\top y \le \|x\|_1 \|y\|_\infty \\ &\le \mathbf{E} \Big\| \sum_{i=1}^n \Omega_i x_i \Big\|_\infty \end{aligned}$$

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

$$\Big\| \sum_{i=1}^n \Omega_i x_i \Big\|_\infty = \max_{j \in 1:d} \Big| \sum_{i=1}^n \Omega_i x_{i,j} \Big| = \max_{j \in 1:d} \Big| \sum_{i=1}^n \Omega_i t_{j,i} \Big| = \max_{t \in \mathcal{T}} \Big| \sum_{i=1}^n \Omega_i t_i \Big|,$$

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 $\omega_1, \dots, \omega_n \in \{-1, 1\}^n$ we have

$$\max_{t \in \mathcal{T}} \Big| \sum_{i=1}^n \omega_i t_i \Big| = \max_{t \in \mathcal{T} \cup \mathcal{T}_-} \sum_{i=1}^n \omega_i t_i,$$

where we have defined $\mathcal{T}_- = \{-t_1, \dots, -t_d\}$, with $-t_j = (-x_{1,j}, \dots, -x_{n,j})$. Hence, we have

$$\mathtt{Rad}(\mathcal{A}_1 \circ \{x_1, \dots, x_n\}) \le \mathtt{Rad}(\mathcal{T} \cup \mathcal{T}_-),$$

and the proof follows by Massart's lemma as

$$\mathtt{Rad}(\mathcal{T} \cup \mathcal{T}_-) \le \max_{t \in \mathcal{T} \cup \mathcal{T}_-} \|t\|_2 \frac{\sqrt{2 \log |\mathcal{T} \cup \mathcal{T}_-|}}{n} \le \sqrt{n} \max_i \|x_i\|_\infty \frac{\sqrt{2 \log(2d)}}{n}.$$

If $c \neq 1$, then we can rescale $w^\top x = (\frac{w}{\|w\|_1})^\top (\|w\|_1 x)$ and the following equivalence concludes the proof:

$$\{x \in \mathbb{R}^d \to w^\top x : w \in \mathbb{R}^d, \|w\|_1 \le c\} = \{x \in \mathbb{R}^d \to w^\top (cx) : w \in \mathbb{R}^d, \|w\|_1 \le 1\}.$$

■

## 3.4 Linear predictors *simplex*/$\ell_\infty$ constraints (Boosting)

**Proposition 3.4** *Let* $\Delta_d := \{w \in \mathbb{R}^d : \|w\|_1 = 1, w_1, \dots, w_d \ge 0\}$ *and let* $\mathcal{A}_\Delta := \{x \in \mathbb{R}^d \to w^\top x : w \in \Delta_d\}$. *Then, for any* $x_1 \dots, x_n \in \mathbb{R}^d$,

$$\boxed{\mathtt{Rad}(\mathcal{A}_\Delta \circ \{x_1, \dots, x_n\}) \le \max_i \|x_i\|_\infty \frac{\sqrt{2 \log d}}{\sqrt{n}}}$$

**Proof:** We have

$$n\,\mathtt{Rad}(\mathcal{A}_\Delta \circ \{x_1, \dots, x_n\}) = \mathbf{E} \sup_{w \in \Delta_d} \sum_{i=1}^n \Omega_i w^\top x_i = \mathbf{E} \sup_{w \in \Delta_d} w^\top \Big( \sum_{i=1}^n \Omega_i x_i \Big).$$

Note that for any vector $v = (v_1, \dots, v_d) \in \mathbb{R}^d$ we have (c.f. Proposition 8.6 for a general statement involving convex hulls and for the proof)

$$\sup_{w \in \Delta_d} w^\top v = \max_{j \in 1:d} v_j.$$

Then,

$$\mathbf{E} \sup_{w \in \Delta_d} w^\top \Big( \sum_{i=1}^n \Omega_i x_i \Big) = \mathbf{E} \max_{j \in 1:d} \sum_{i=1}^n \Omega_i x_{i,j} = n \, \mathtt{Rad}(\mathcal{T}),$$

with $\mathcal{T} = \{t_1 \dots, t_d\}$ with $t_j = (x_{1,j}, \dots, x_{n,j})$ for any $j \in \{1, \dots, d\}$. The proof follows by Massart's lemma as

$$\mathtt{Rad}(\mathcal{T}) \le \max_{t \in \mathcal{T}} \|t\|_2 \frac{\sqrt{2 \log |\mathcal{T}|}}{n} \le \sqrt{n} \max_i \|x_i\|_\infty \frac{\sqrt{2 \log d}}{n}.$$

■

## 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 $l^{(k)} : \mathbb{R}^{d_{k-1}} \to \mathbb{R}^{d_k}$ consists of a coordinate-wise composition of an activation function $\sigma^{(k)} : \mathbb{R} \to \mathbb{R}$ and an affine map, namely,

$$l^{(k)}(x) := \sigma^{(k)}(\mathbf{w}^{(k)} x + b^{(k)}),$$

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

A feed-forward neural network with depth $\iota$ (and $\iota - 1$ hidden layers) is given by the function $f^\iota_{nn} : \mathbb{R}^d \to \mathbb{R}$ defined as

$$f^{(\iota)}_{nn}(x) := l^{(\iota)} \circ \cdots \circ l^{(1)}(x) \equiv l^{(\iota)}(\cdots l^{(2)}(l^{(1)}(x)) \cdots),$$

with $d_0 = d$, $d_\iota = 1$, $\sigma^{(r)} = \sigma$ for a given function $\sigma$ for all $r < \iota$, and $\sigma^{(\iota)}(x) = x$ (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

$$\mathcal{A}^{(\iota)}_{nn} := \{x \in \mathbb{R}^d \to f^{(\iota)}_{nn}(x) : \|\mathbf{w}^{(k)}\|_\infty \le \omega, \|b^{(k)}\|_\infty \le \beta \ \forall k\}, \tag{3.1}$$

where for a given matrix $\mathbf{m}$, the $\ell_\infty$ norm is defined as $\|\mathbf{m}\|_\infty := \max_i \sum_j |\mathbf{m}_{ij}|$.

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 3.5** *Let $\mathcal{L}$ be a class of functions from $\mathbb{R}^d$ to $\mathbb{R}$ that includes the zero function. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be $\gamma$-Lipschitz and define $\mathcal{L}' := \{x \in \mathbb{R}^d \to \sigma(\sum_{j=1}^m w_j l_j(x) + b) \in \mathbb{R} : |b| \le \beta, \|w\|_1 \le \omega, l_1, \dots, l_m \in \mathcal{L}\}$. Then, for any $x_1, \dots, x_n \in \mathbb{R}^d$,*

$$\boxed{\mathtt{Rad}(\mathcal{L}' \circ \{x_1, \dots, x_n\}) \le \gamma \Big( \frac{\beta}{\sqrt{n}} + 2 \, \omega \, \mathtt{Rad}(\mathcal{L} \circ \{x_1, \dots, x_n\}) \Big)} \tag{3.2}$$

**Proof:** We give a proof that makes use of several properties of Rademacher complexities introduced in the previous lecture. Let

$$\mathcal{F} := \{x \in \mathbb{R}^d \to \sum_{i=1}^{m} w_j l_j(x) \in \mathbb{R} : \|w\|_1 \leq \omega, l_1, \ldots, l_m \in \mathcal{L}\},$$
$$\mathcal{G} := \{x \in \mathbb{R}^d \to b \in \mathbb{R} \ : |b| \leq \beta\}.$$

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

$$\texttt{Rad}(\mathcal{L}' \circ \{x_1, \ldots, x_n\}) \leq \gamma\Big(\texttt{Rad}(\mathcal{F} \circ \{x_1, \ldots, x_n\}) + \texttt{Rad}(\mathcal{G} \circ \{x_1, \ldots, x_n\})\Big).$$

We first address the term $\texttt{Rad}(\mathcal{F} \circ \{x_1, \ldots, x_n\}$. As $\mathcal{L}$ contains the zero function by assumption, we will show that

$$\texttt{Rad}(\mathcal{F} \circ \{x_1, \ldots, x_n\}) \leq \omega\, \texttt{Rad}(\text{conv}(\mathcal{L} - \mathcal{L}) \circ \{x_1, \ldots, x_n\}),$$

where $\mathcal{L} - \mathcal{L} = \{l - l' : l \in \mathcal{L}, l' \in \mathcal{L}\}$. First of all, note that

$$\texttt{Rad}(\mathcal{F} \circ \{x_1, \ldots, x_n\} = \texttt{Rad}(\mathcal{F}' \circ \{x_1, \ldots, x_n\})$$

where

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

(this is because the maximum of a linear function of $w$ over the constraint $\|w\|_1 \leq \omega$ is achieved for the values satisfying $\|w\|_1 = \omega$; we already saw this type of arguments in the proof of Proposition 2.8). Then, note that for any $w \in \mathbb{R}^m$ such that $\|w\|_1 = 1$ we have

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

(here 0 represents the zero function) which is a convex combination of elements in $\mathcal{L} - \mathcal{L}$. Hence, by applying in order the convex hall property, the summation property, and the scalar multiplication property of Rademacher complexities, we find

$$\begin{aligned}\texttt{Rad}(\mathcal{F} \circ \{x_1, \ldots, x_n\}) &\leq \omega\, \texttt{Rad}(\text{conv}(\mathcal{L} - \mathcal{L}) \circ \{x_1, \ldots, x_n\}) = \omega\, \texttt{Rad}((\mathcal{L} - \mathcal{L}) \circ \{x_1, \ldots, x_n\}) \\ &= \omega\, \texttt{Rad}(\mathcal{L} \circ \{x_1, \ldots, x_n\}) + \omega\, \texttt{Rad}(-\mathcal{L} \circ \{x_1, \ldots, x_n\}) = 2\,\omega\, \texttt{Rad}(\mathcal{L} \circ \{x_1, \ldots, x_n\}).\end{aligned}$$

We now address the term $\texttt{Rad}(\mathcal{G} \circ \{x_1, \ldots, x_n\})$. We have

$$n\, \texttt{Rad}(\mathcal{G} \circ \{x_1, \ldots, x_n\})\Big) = \mathbf{E} \sup_{b: |b| \leq \beta} b \sum_{i=1}^{n} \Omega_i \leq \mathbf{E} \sup_{b: |b| \leq \beta} |b| \Big| \sum_{i=1}^{n} \Omega_i \Big| = \beta\, \mathbf{E} \Big| \sum_{i=1}^{n} \Omega_i \Big| \leq \beta \sqrt{n},$$

where the last inequality follows by Jensen's inequality, as $\mathbf{E}|\sum_{i=1}^n \Omega_i| = \mathbf{E}\sqrt{(\sum_{i=1}^n \Omega_i)^2} \leq \sqrt{\mathbf{E}[(\sum_{i=1}^n \Omega_i)^2]} = \sqrt{n}$ 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 3.5 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 3.3.

**Proposition 3.6** *Let $\sigma$ be $\lambda$-Lipschitz. Let $\mathcal{A}_{nn}^{(\iota)}$ be defined as in 3.1. Then, for any $x_1 \ldots, x_n \in \mathbb{R}^d$,*

$$\boxed{\texttt{Rad}(\mathcal{A}_{nn}^{(\iota)} \circ \{x_1, \ldots, x_n\}) \leq \frac{1}{\sqrt{n}}\Big(\beta + 2\omega\beta\lambda \sum_{k=0}^{\iota-3} (2\omega\lambda)^k + 2\omega(2\omega\lambda)^{\iota-2} \max_i \|x_i\|_\infty \sqrt{2\log(2d)}\Big)}$$

**Proof:** As the last layer of the neural network is linear, i.e., $\sigma^{(\iota)}(x) = x$, we can apply Proposition 3.5 with $\gamma = 1$ (as $\sigma^{(\iota)}$ is 1-Lipschitz) once and then apply (3.2) in Proposition 3.5 with $\gamma = \lambda$ for $\iota - 2$ times. We find

$$\texttt{Rad}(\mathcal{A}_{nn}^{(\iota)} \circ \{x_1, \ldots, x_n\}) \leq \frac{\beta}{\sqrt{n}} + 2\omega\Big(\frac{\beta\lambda}{\sqrt{n}} \sum_{k=0}^{\iota-3} (2\omega\lambda)^k + (2\omega\lambda)^{\iota-2} \texttt{Rad}(\mathcal{A}_1 \circ \{x_1, \ldots, x_n\})\Big).$$

■

## 3.6 Remarks on generalization errors

The bounds derived in this lecture assert that to achieve good generalization guarantees one needs to use a training sample size $n$ that exceeds the "complexity" of the learning model, where complexity is captured by certain functions of the *number* of model parameters, their *magnitude*, and possibly other structural properties (Lipschitz constants, etc.). In the case of linear models, the number of model parameters $p$ coincides with the dimensionality of the data $d$, i.e. $p = d$. In the examples we considered in this lecture, the complexity of these models is controlled by either the square-root of the number of parameters, $\sqrt{p}$, for $\ell_2/\ell_2$ constraints, or by the square-root of the logarithm of the number of parameters, $\sqrt{\log p}$, for $\ell_1/\ell_\infty$ and *simplex*/$\ell_\infty$ constraints. The magnitude of the model parameters plays a role as multiplicative factors depending on either the $\ell_2$ or $\ell_1$ norms. In the example of a feed-forward neural network, on the other hand, the model parameters are $\{\mathbf{w}^{(k)}, b^{(k)}, k \in \{1, \ldots, \iota\}\}$, so in general $p \neq d$ and typically $p \gg d$. Here the interplay between the number of model parameters, their magnitude (as captures by the boundedness constants $\omega$ and $\beta$) and other structural properties (the Lipschitz constant $\lambda$) is more involved, as already attested by the bound given in Proposition 3.6 (note that for this bound not to become trivial in the limit of an infinite number of layers, $\iota \to \infty$, it would have to be $2\omega\lambda < 1$, which is a restrictive requirement not needed in practice). Investigating meaningful notions of complexity for neural networks is an active topic of research.