

Algorithmic Foundations of Learning

Lecture 3 Rademacher complexity. Examples

Patrick Rebeschini

Department of Statistics
University of Oxford

Recap

► Goal:

$$\mathbf{E} \underbrace{r(A^*) - r(a^*)}_{\text{estimation error for ERM}} \lesssim \frac{f(\text{dimension})}{n^\alpha}$$

► Sufficient:

$$\mathbf{E} \sup_{a \in \mathcal{A}} \{r(a) - R(a)\} \leq \frac{f(\text{dimension}, \text{complexity of } \mathcal{A})}{n^\alpha}$$

Bound in expectation (Proposition 2.3)

If the loss function ℓ is bounded by c , we have

$$\mathbf{E} \max_{a \in \mathcal{A}} \{r(a) - R(a)\} \leq c \frac{\sqrt{2 \log |\mathcal{A}|}}{\sqrt{n}}$$

Bound in expectation via Rademacher complexity (Proposition 2.11)

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

with $\mathcal{L} \circ \{Z_1, \dots, Z_n\} := \{(\ell(a, Z_1), \dots, \ell(a, Z_n)) \in \mathbb{R}^n : a \in \mathcal{A}\}$

Note

If $|\mathcal{A}| < \infty$, Massart's Lemma recovers previous result (modulo constant)

Massart's Lemma (Lemma 2.9)

Let $\mathcal{T} \subseteq \mathbb{R}^n$ and $\bar{t} := \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} t$. We have

$$\text{Rad}(\mathcal{T}) \leq \max_{t \in \mathcal{T}} \|t - \bar{t}\|_2 \frac{\sqrt{2 \log |\mathcal{T}|}}{n}$$

We have

- ▶ $\mathcal{T} = \mathcal{L} \circ \{Z_1, \dots, Z_n\} = \{(\ell(a, Z_1), \dots, \ell(a, Z_n)) \in \mathbb{R}^n : a \in \mathcal{A}\}$
- ▶ $|\mathcal{T}| \leq |\mathcal{A}|$
- ▶ $\|t - \bar{t}\|_2 \leq \|t\|_2 = \sqrt{\sum_{i=1}^n \ell(a, Z_i)^2} \leq \sqrt{nc^2} = c\sqrt{n}$

so we obtain

$$\mathbf{E} \max_{a \in \mathcal{A}} \{r(a) - R(a)\} \leq 2c \frac{\sqrt{2 \log |\mathcal{A}|}}{\sqrt{n}}$$

For the proof, let's review some basic properties of conditional expectations...

Properties of conditional expectations

Let X, Y be real-valued random variables. The following can be made precise:

- ▶ $\mathbf{E}X$ is the “best” estimate of X with no information. It is a **constant**
- ▶ $\mathbf{E}[X|Y]$ is the “best” estimate of X if we know Y . It is a **random variable**
- ▶ If X and Y are independent, Y does not contain any information on X and

$$\mathbf{E}[X|Y] = \mathbf{E}X \quad \text{independence property (a)}$$

- ▶ If f is a deterministic function, if we know Y we also know $f(Y)$ and

$$\mathbf{E}[Xf(Y)|Y] = f(Y)\mathbf{E}[X|Y] \quad \text{“taking out what is known” property (b)}$$

- ▶ Law of total expectation (“Ignorants win in life” phenomenon)

$$\mathbf{E}\mathbf{E}[X|Y] = \mathbf{E}X \quad \text{“tower” property (c)}$$

Remark: the above holds with $\mathbf{E} \rightarrow \mathbf{E}[\cdot|Z]$, $\mathbf{E}[\cdot|Y] \rightarrow \mathbf{E}[\cdot|Y, Z]$ possibly using the notion of conditional independence.

Proof: Symmetrization

Proof

- $S = \{Z_1, \dots, Z_n\}$ and $\tilde{S} = \{\tilde{Z}_1, \dots, \tilde{Z}_n\}$ be independent samples with same distribution

$$r(a) = \mathbf{E} \ell(a, Z) = \frac{1}{n} \sum_{i=1}^n \mathbf{E} \ell(a, \tilde{Z}_i) \stackrel{(a)}{=} \frac{1}{n} \sum_{i=1}^n \mathbf{E}[\ell(a, \tilde{Z}_i) | S]$$

- By properties of conditional expectations (tower property and others) we get

$$\begin{aligned} \mathbf{E} \sup_{a \in \mathcal{A}} \{r(a) - R(a)\} &= \mathbf{E} \sup_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^n \left(\mathbf{E}[\ell(a, \tilde{Z}_i) | S] - \ell(a, Z_i) \right) \\ &\stackrel{(b)}{=} \mathbf{E} \sup_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^n \mathbf{E}[\ell(a, \tilde{Z}_i) - \ell(a, Z_i) | S] \\ &\leq \mathbf{E} \mathbf{E} \left[\sup_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^n \{\ell(a, \tilde{Z}_i) - \ell(a, Z_i)\} \middle| S \right] \\ &\stackrel{(c)}{=} \mathbf{E} \sup_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^n \{\ell(a, \tilde{Z}_i) - \ell(a, Z_i)\} \\ &= \mathbf{E} \sup_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^n \Omega_i \{\ell(a, \tilde{Z}_i) - \ell(a, Z_i)\} \\ &\leq 2 \mathbf{E} \sup_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^n \Omega_i \ell(a, Z_i) = 2 \mathbf{E} \text{Rad}(\mathcal{L} \circ \{Z_1, \dots, Z_n\}) \end{aligned}$$

Supervised Learning. Regression

Today, we consider the setting of regression:

- ▶ $Z_i = (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$
- ▶ Admissible action set $\mathcal{A} \subseteq \mathcal{B} := \{a : \mathbb{R}^d \rightarrow \mathbb{R}\}$
- ▶ Loss function is of the form $\ell(a, (x, y)) = \phi(a(x), y)$, for $\phi : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_+$

(Proposition 3.1)

If the function $\hat{y} \rightarrow \phi(\hat{y}, y)$ is γ -Lipschitz for any $y \in \mathcal{Y}$, then

$$\text{Rad}(\mathcal{L} \circ \{z_1, \dots, z_n\}) \leq \gamma \text{Rad}(\mathcal{A} \circ \{x_1, \dots, x_n\})$$

with $\mathcal{A} \circ \{x_1, \dots, x_n\} := \{(a(x_1), \dots, a(x_n)) \in \mathbb{R}^n : a \in \mathcal{A}\}$

- ▶ New goal:
$$\text{Rad}(\mathcal{A} \circ \{x_1, \dots, x_n\}) \leq \frac{f(\text{dimension, complexity of } \mathcal{A})}{n^\alpha}$$

Linear predictors ℓ_2/ℓ_2 constraints (SVM)

(Proposition 3.2)

Let $\mathcal{A}_2 := \{x \in \mathbb{R}^d \rightarrow w^\top x : w \in \mathbb{R}^d, \|w\|_2 \leq c\}$. Then,

$$\text{Rad}(\mathcal{A}_2 \circ \{x_1, \dots, x_n\}) \leq c \frac{\max_i \|x_i\|_2}{\sqrt{n}}$$

Note: typically, $\max_i \|x_i\|_2 \sim \sqrt{d}$ as

$$\|x\|_2 = \sqrt{\sum_{i=1}^d x_i^2} \leq \sqrt{d} \max_{i \in [d]} |x_i|$$

Proof

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

Linear predictors *simplex*/ ℓ_∞ constraints (Boosting)

Define d -dimensional simplex: $\Delta_d := \{w \in \mathbb{R}^d : \|w\|_1 = 1, w_1, \dots, w_d \geq 0\}$.

(Proposition 3.4)

Let $\mathcal{A}_\Delta := \{x \in \mathbb{R}^d \rightarrow w^\top x : w \in c\Delta_d\}$. Then

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

Note: typically, $\max_i \|x_i\|_\infty \not\propto d$, so overall dependence is $\sim \sqrt{\log d}$

(Similar result for Proposition 3.3 for ℓ_1/ℓ_∞ constraints. In that case we present a different argument in the lecture notes, based on Hölder's inequality $x^\top y \leq \|x\|_1 \|y\|_\infty$. The same argument used for the *simplex*/ ℓ_∞ case also works)

Remark: Difference between d and $\log d$ is ultimately linked with the different dependence with the dimension d for the ℓ_2 and ℓ_1 ball, respectively.

Proof

- ▶ We have

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

- ▶ Note that for any vector $v = (v_1, \dots, v_d) \in \mathbb{R}^d$ we have

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

- ▶ Then,

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

with $\mathcal{T} = \{t_1, \dots, t_d\}$ with $t_j = (x_{1,j}, \dots, x_{d,j})$ for any $j \in \{1, \dots, d\}$

- ▶ The proof follows by Massart's lemma as

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

Feed-forward neural networks

- ▶ A **layer** $l^{(k)} : \mathbb{R}^{d_{k-1}} \rightarrow \mathbb{R}^{d_k}$ consists of a coordinate-wise composition of an activation function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ and an affine map:

$$l^{(k)}(x) := \sigma(W^{(k)}x + b^{(k)})$$

- ▶ A **neural network** with **depth** ℓ is the function (with $d_0 = d, d_\ell = 1$)

$$f_{nn}^\ell : x \in \mathbb{R}^d \longrightarrow f_{nn}^\ell(x) := l^{(\ell)}(\dots l^{(2)}(l^{(1)}(x))\dots)$$

(Proposition 3.6)

Let $\mathcal{A}_{nn}^{(\ell)} := \{x \in \mathbb{R}^d \rightarrow f_{nn}^\ell(x) : \|\mathbf{w}^{(k)}\|_\infty \leq \omega, \|b^{(k)}\|_\infty \leq \beta \forall k\}$.

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