**Corollary 17.1.** $\forall p \in \mathcal{P}$ , $\forall n$ , $\forall \widehat{f}$ , $\mathcal{F} : Z \longmapsto [0,1]$

*(1)* $\mathcal{R}(\widehat{f}) \leq \widehat{\mathcal{R}}(\widehat{f}) + \mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{\log 1/\delta}{2n}}$ *with probability at least* $1-\delta$

*(2)* $\mathcal{R}(\widehat{f}_{ERM}) \leq \mathcal{R}(f^*) + 2\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{2\log 1/\delta}{n}}$

**Proof:** By McDiarmid inequality:

$$P(\Delta_n(\mathcal{F}) - \mathbb{E}[\Delta_n(\mathcal{F})] \geq t) \leq e^{-2nt^2} \tag{17.1}$$

So with probability at least $1-\delta$, $t = \sqrt{\frac{\log 1/\delta}{2n}}$:

$$\Delta_n(\mathcal{F}) \leq \mathbb{E}[\Delta_n(\mathcal{F})] + \sqrt{\frac{\log 1/\delta}{2n}} \tag{17.2}$$

Then using the theorem in the previous lecture we get part (1). Also the following gives part (2):

$$\mathcal{R}(\widehat{f}) \leq \mathcal{R}(f^*) + 2\left(\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{\log 1/\delta}{2n}}\right) \tag{17.3}$$

## 17.1 Properties of $\mathcal{R}_n(\mathcal{F})$

**Theorem 17.2** (Bartlett and Mendelson (2003))**.** *Let* $\mathcal{F}, \mathcal{F}_1, ..., \mathcal{F}_k, \mathcal{H}$ *be classes of functions:*

*(1)* *If* $\mathcal{F} \subseteq \mathcal{H}$ *then* $\mathcal{R}_n(\mathcal{F}) \leq \mathcal{R}_n(\mathcal{H})$

*(2)* $\mathcal{R}_n(\mathcal{F}) = \mathcal{R}_n(conv\ \mathcal{F}) = \mathcal{R}_n(abs\ conv\ \mathcal{F})$

*(3)* $\mathcal{R}_n(c\mathcal{F}) = |c| \cdot \mathcal{R}_n(\mathcal{F})$

*(4)* $\phi : \mathbb{R} \longmapsto \mathbb{R}$ *and* $\phi$ *is L-Lipshcitz and* $\phi(0) = 0$ *then* $\mathcal{R}_n(\phi \circ \mathcal{F}) \leq 2L\mathcal{R}_n(\mathcal{F})$

*(5)* *For any uniformely bounded* $h$ *we have* $\mathcal{R}_n(\mathcal{F} + h) \leq \mathcal{R}_n(\mathcal{F}) + ||h||_\infty/\sqrt{n}$

*(6)* $1 \leq q \leq \infty$ *define* $\mathcal{L}(\mathcal{F}, h, q) = \{|f-h|^q : f \in \mathcal{F}\}$*, if* $||\mathcal{F} - h||_\infty \leq 1$ *,* $\forall f$ *then* $\mathcal{R}_n(\mathcal{L}(\mathcal{F}, h, q)) \leq 2q\big(\mathcal{R}_n(\mathcal{F}) + ||h||_\infty/\sqrt{n}\big)$

*(7)* $\mathcal{R}_n(\sum_{i=1}^k \mathcal{F}_i) \leq \sum_{i=1}^k \mathcal{R}_n(\mathcal{F}_i)$

## 17.2 Examples

**Lemma 17.3.** *Let $x \in \mathbb{R}^p$, define $\mathcal{F} = \{x \longmapsto \langle x, w \rangle, w \in \mathbb{R}^p, ||w||_1 \leq 1\}$. $\forall x_1, ..., x_n \in \mathbb{R}^p$: $\widehat{\mathcal{R}}_n(\mathcal{F}) \leq \frac{2}{n} \max_j ||x_j||_2 \sqrt{2 \log p}$.*

**Proof:**

$$\begin{aligned}
\mathbb{E}_\sigma \left[ \sup_{f \in \mathcal{F}} \frac{2}{n} \sum \sigma_i f(x_i) \right] &= \mathbb{E}_\sigma \left[ \sup_{||w||_1 \leq 1} \frac{2}{n} \sum \sigma_i \langle w, x_i \rangle \right] \\
&= \mathbb{E}_\sigma \left[ \sup_{||w||_1 \leq 1} \langle w, \frac{2}{n} \sum \sigma_i x_i \rangle \right] \\
&= \mathbb{E}_\sigma \left[ \max_j \frac{2}{n} \sum_i \sigma_i x_{ij} \right] \\
&= \frac{2}{n} \mathbb{E}_\sigma \left[ \max_j Z_j \right] \ , \ Z_j = \sum_i \sigma_i x_{ij}
\end{aligned} \tag{17.4}$$

Now we have:

$$\mathbb{E}_\sigma \left[ e^{\lambda Z_j} \right] = \mathbb{E}_\sigma \left[ e^{\lambda \sum_i \sigma_i x_{ij}} \right] = \prod_{i=1}^n \mathbb{E}_\sigma \left[ e^{\lambda \sigma_i x_{ij}} \right] \leq \prod_{i=1}^n e^{\lambda^2 x_{ij}^2 / 2} = e^{\lambda^2 ||x_j||_2^2 / 2} \tag{17.5}$$

and so $Z_j$ is $||x_j||_2$ sub-Gaussian and so:

$$\mathbb{E} \max_j Z_j \leq \log \left( \sum_{j=1}^p e^{\lambda^2 ||x_j||_2^2 / 2} \right) \leq \log \left( p e^{\lambda^2 \max_j ||x_j||_2^2 / 2} \right) \tag{17.6}$$

Now from Lecture 12, we get:

$$\widehat{\mathcal{R}}(\mathcal{F}) \leq \frac{2 \max_j ||x_j||_2}{n} \sqrt{2 \log p} \tag{17.7}$$

### 17.2.1 Neural networks

**Theorem 17.4.** *Suppose $\sigma : \mathbb{R} \longmapsto [-1, 1]$ is the activation function and is L-Lipshcitz and $\sigma(0) = 0$. Define $\mathcal{F}$ to be a 2-layer neural network with 1-norm constraints on the weights: $\mathcal{F} = \{x \longmapsto \sum_i w_i \sigma(\langle v_i, x \rangle) : ||w||_1 \leq 1, v_i \leq B\}$. Then for inputs $x_1, ...x_n \in \mathbb{R}^p$: $\widehat{\mathcal{R}}(\mathcal{F}) \leq \frac{2LB}{n} \max_j ||x_j||_2 \sqrt{2 \log p}$.*

### 17.2.2 Kernel methods

A side on kernel methods:

1. To every kernel $k : \mathcal{X} \times \mathcal{X} \longmapsto \mathbb{R}$, we can associate a feature map $\Phi : \mathcal{X} \longmapsto \mathcal{H}$, where $\mathcal{H}$ is the Hilbert space with inner product $\langle ., . \rangle$ and $\forall x_1, x_2 \in \mathcal{X}$: $k(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle$.

2. If $||.||$ is the norm on $\mathcal{H}$, then $||\sum \alpha_2 \Phi(x_i)||^2 = \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j)$.

**Theorem 17.5.** *For $x_a, ..., x_n$ random elements of $\mathcal{X}$, let $l : \mathcal{Y} \times \mathbb{R} \longmapsto [0,1]$ be L-Lipshcitz with $l(0) = 0$. Define $\mathcal{F} = \{x \longmapsto \sum \alpha_i k(x, x_i), \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) \leq B^2\} \subseteq \{x \longmapsto \langle w, \Phi(x) \rangle : ||w|| \leq B\}$, then $\mathcal{R}_n(l \circ \mathcal{F}) \leq 4BL\sqrt{\frac{\mathbb{E}k(x,x)}{n}}$.*

# References


Bartlett, P.L., and Mendelson, S. (2003), "Rademacher and Gaussian Complexities: Risk Bounds and Structural Results," *The Journal of Machine Learning Research*, **3**, 463–482.