Algorithmic Foundations of Learning **Lecture 2**

Maximal Inequalities and Rademacher Complexity

Lecturer: Patrick Rebeschini *Version: October 15th 2019*

2.1 Introduction

In this lecture we develop tools to bound the supremum of empirical processes. Recall the setting of prediction in statistical learning. Let $S = \{Z_1, \dots, Z_n\} \in \mathcal{Z}^n$ be a family of independent random variables on $\mathcal{Z}$ drawn from a certain distribution, and let Z be an independent sample from the same distribution. Let $\mathcal{A}$ be a given set of actions (a.k.a. decision rules in the general setting; or predictors/classifiers in the setting of supervised learning), and $\ell : \mathcal{A} \times \mathcal{Z} \to \mathbb{R}_+$ be a given loss function. Let $r(a) := \mathbf{E}\,\ell(a, Z)$ be the population risk, and $R(a) := \frac{1}{n}\sum_{i=1}^n \ell(a, Z_i)$ be the empirical risk. We aim to develop tools to produce bounds of the following type:

$$\mathbf{E} \sup_{a \in \mathcal{A}} \{r(a) - R(a)\} \leq \boxed{\ ?\ }$$

where $\boxed{\ ?\ }$ is a function of the number of data points n and certain notions of complexity for the class $\mathcal{A}$. As for any fixed $a \in \mathcal{A}$ we have $\mathbf{E}R(a) = \mathbf{E}\,\ell(a, Z) = r(a)$, recall that the Law of Large Numbers (LLN) corresponds to the statement

$$r(a) - R(a) \to 0 \text{ as } n \to \infty,$$

where the limit is taken almost surely for the Strong LLN, or in probability for the Weak LLN. In either cases, the LLN implies convergence in distribution, i.e.,

$$\mathbf{E}\{r(a) - R(a)\} \to 0 \text{ as } n \to \infty.$$

Therefore, the type of bounds that want to derive corresponds to a non-asymptotic (as n is finite, not going to infinity), uniform (as we consider the supremum over all $a \in \mathcal{A}$, not a fixed a) LLN in expectation. If $\mathcal{A}$ is finite, then its cardinality $|\mathcal{A}|$ is the notion of complexity we look for (Proposition 2.3). In the general case when $\mathcal{A}$ is infinite, the notion of complexity is given by the *Rademacher complexity* (Proposition 2.11), which we define and discuss below.

2.2 Hoeffding's Lemma

Throughout this course, we will use repeatedly the following lemma that bounds the moment generating function of bounded random variables.

Lemma 2.1 (Hoeffding) *Let X be a real-valued random variable such that $a \leq X \leq b$. Then, for any $\lambda \in \mathbb{R}$ we have*

$$\boxed{\mathbf{E}\, e^{\lambda(X - \mathbf{E}X)} \leq e^{\lambda^2 (b-a)^2/8}}$$

Proof: Without loss of generality, consider a random variable with zero mean, namely, $\mathbf{E}X = 0$. Let

$\psi(\lambda) = \log \mathbf{E}\, e^{\lambda X}$. The first and second derivatives of ψ read, respectively,

$$\psi'(\lambda) = \frac{\mathbf{E}[Xe^{\lambda X}]}{\mathbf{E}\, e^{\lambda X}},$$
$$\psi''(\lambda) = \frac{\mathbf{E}[X^2 e^{\lambda X}]}{\mathbf{E}\, e^{\lambda X}} - \left(\frac{\mathbf{E}[Xe^{\lambda X}]}{\mathbf{E}\, e^{\lambda X}}\right)^2.$$

We can interpret $\psi''(\lambda)$ as the variance of the random variable X under the tilted probability distribution $\mathbf{Q}(\mathrm{d}x) = \frac{e^{\lambda X}}{\mathbf{E}\, e^{\lambda X}}\mathbf{P}(\mathrm{d}x)$. As the variance does not change upon translations by constants, we get

$$\psi''(\lambda) = \mathbf{E}_{\mathbf{Q}}[X^2] - (\mathbf{E}_{\mathbf{Q}}X)^2 = \mathbf{Var}_{\mathbf{Q}}X = \mathbf{Var}_{\mathbf{Q}}\Big(X - \frac{a+b}{2}\Big) \le \mathbf{E}_{\mathbf{Q}}\Big[\Big(X - \frac{a+b}{2}\Big)^2\Big] \le \frac{(b-a)^2}{4}.$$

By the Fundamental Theorem of Calculus we have

$$\psi(\lambda) = \int_0^\lambda \int_0^\mu \psi''(\rho)\mathrm{d}\rho\mathrm{d}\mu \le \frac{\lambda^2(b-a)^2}{8}.$$

■

2.3 Maximum of finitely many random variables

The maximum of a collection of finitely-many bounded random variables (not necessarily independent) growths logarithmically in the number of random variables. This follows as an application of Hoeffding's lemma.

Proposition 2.2 *Let $X_1, \ldots, X_n$ be n centered random variables (i.e., $\mathbf{E}X_i = 0$) bounded in the interval $[a, b]$. Then,*

$$\boxed{\mathbf{E}\max_{i\in[n]} X_i \le \frac{b-a}{\sqrt{2}}\sqrt{\log n}}$$

Proof: We adopt two standard techniques in probability to prove upper bounds: first, we take exponentials and use Jensen's inequality; second, we bound the maximum of a set of non-negative numbers by its sum. For the first step, note that for any real-valued random variable X and any $\lambda > 0$, Jensen's inequality yields

$$\mathbf{E}X = \frac{1}{\lambda}\log e^{\lambda \mathbf{E}X} \le \frac{1}{\lambda}\log \mathbf{E}\, e^{\lambda X},$$

as the function $x \to e^{\lambda x}$ is convex. For the second step, note that if $X = \max_{i\in[n]} X_i$, then

$$\mathbf{E}\, e^{\lambda X} = \mathbf{E}\, e^{\lambda \max_{i\in[n]} X_i} = \mathbf{E}\max_{i\in[n]} e^{\lambda X_i} \le \mathbf{E}\sum_{i=1}^n e^{\lambda X_i} = \sum_{i=1}^n \mathbf{E}\, e^{\lambda X_i}.$$

By Hoeffding's lemma, Lemma 2.1, we have

$$\mathbf{E}\, e^{\lambda X_i} \le e^{\lambda^2 (b-a)^2/8},$$

and the above yields

$$\mathbf{E}\max_{i\in[n]} X_i \le \frac{1}{\lambda}\log\sum_{i=1}^n e^{\lambda^2(b-a)^2/8} = \frac{1}{\lambda}\log n + \frac{\lambda(b-a)^2}{8}.$$

The bound is of the form $\alpha/\lambda + \lambda\beta$, for $\alpha = \log n$ and $\beta = (b-a)^2/8$. Optimizing this bound over $\lambda > 0$, the maximum is at $\lambda = \sqrt{\alpha/\beta} = \sqrt{8\log n/(b-a)^2}$ and yields the optimal value $2\sqrt{\alpha\beta} = (b-a)\sqrt{\log n/2}$. ■

2.4 Maximum of empirical processes

The same idea yields an upper bound for the quantity $\mathbf{E}\max_{a\in\mathcal{A}}\{r(a)-R(a)\}$, where we now exploit the independence of the random variables $Z_1,\ldots,Z_n$ before applying Hoeffding's lemma. This bound is non-trivial when $|\mathcal{A}|<\infty$.

Proposition 2.3 *Let the loss function ℓ be uniformly bounded by 1, i.e., $0\le\ell(a,z)\le 1$ for any $a\in\mathcal{A}$ and $z\in\mathcal{Z}$. Then,*

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

Proof: We proceed as in the proof of Proposition 2.2 with the substitutions $[n]\to\mathcal{A}$ and $i\to a$, choosing $X_a=\frac{1}{n}\sum_{i=1}^n\{r(a)-\ell(a,Z_i)\}$ and $X=\max_{a\in\mathcal{A}}X_a=\max_{a\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^n\{r(a)-\ell(a,Z_i)\}$. By the independence of the random variables $Z_1,\ldots,Z_n$, and by Hoeffding's lemma, Lemma 2.1, as $0\le r(a)=\mathbf{E}\,\ell(a,Z)\le 1$ and $-1\le r(a)-\ell(a,Z_i)\le 1$, we get

$$\mathbf{E}\,e^{\lambda X_a}=\mathbf{E}\,e^{\lambda\sum_{i=1}^n\{r(a)-\ell(a,Z_i)\}/n}=\prod_{i=1}^n\mathbf{E}\,e^{\lambda\{r(a)-\ell(a,Z_i)\}/n}\le\left(e^{\lambda^2(2)^2/(8n^2)}\right)^n=e^{\lambda^2/(2n)},$$

Putting everything together, we get

$$\mathbf{E}\max_{a\in\mathcal{A}}\{r(a)-R(a)\}\le\frac{1}{\lambda}\log\sum_{a\in\mathcal{A}}e^{\lambda^2/(2n)}=\frac{1}{\lambda}\log\left(|\mathcal{A}|e^{\lambda^2/(2n)}\right)=\frac{1}{\lambda}\log|\mathcal{A}|+\frac{\lambda}{2n}.$$

The bound is of the form $\alpha/\lambda+\lambda\beta$, for $\alpha=\log|\mathcal{A}|$ and $\beta=1/(2n)$. Optimizing this bounds over $\lambda>0$, the maximum is at $\lambda=\sqrt{\alpha/\beta}=\sqrt{2n\log|\mathcal{A}|}$ and yields the optimal value $2\sqrt{\alpha\beta}=\sqrt{2\log|\mathcal{A}|/n}$. ■

Remark 2.4 *The condition that the loss function is uniformly bounded by 1 in Proposition 2.3 is without loss of generality. In fact, if the loss function is uniformly bounded by $c>0$, we can rescale the quantity of interest by c, i.e., $c\,\mathbf{E}\sup_{a\in\mathcal{A}}\{r(a)-R(a)\}/c$ and incorporate the division by c into a new loss function that is now upper bounded by 1. The final bound is $c\sqrt{2\log|\mathcal{A}|/n}$.*

In this setting, the cardinality $|\mathcal{A}|$ plays the notion of complexity of the set $\mathcal{A}$. When $|\mathcal{A}|=\infty$ the upper bound in Proposition 2.3 is also infinity (i.e., the upper bound is still true but not useful!). In order to establish nontrivial bounds in the case when the set $\mathcal{A}$ is infinite, we need to replace the cardinality of $\mathcal{A}$ by another notion of complexity of the set $\mathcal{A}$. This notion of complexity is the Rademacher complexity, which we define next along with some of its main properties.

2.5 Rademacher complexity

Henceforth, let $\Omega_1,\ldots,\Omega_n\in\{-1,1\}$ be independent Rademacher random variables (independent of all the other random variables in our model, if any), that are defined as: $\Omega_i=1$ with probability $1/2$ and $\Omega_i=-1$ with probability $1/2$. The following is the definition of the Rademacher complexity for a subset of $\mathbb{R}^n$.

Definition 2.5 *The Rademacher complexity of a set $\mathcal{T}\subseteq\mathbb{R}^n$ is defined as*

$$\boxed{\mathtt{Rad}(\mathcal{T}):=\mathbf{E}\sup_{t\in\mathcal{T}}\frac{1}{n}\sum_{i=1}^n\Omega_i t_i}$$

The quantity $\mathtt{Rad}(\mathcal{T})$ is a measure of complexity of the set $\mathcal{T}$, as $\sup_{t\in\mathcal{T}} \sum_{i=1}^n \Omega_i t_i$ describes how well elements in $\mathcal{T}$ can replicate the sign pattern of a random signal $(\Omega_1, \ldots, \Omega_n) \in \mathbb{R}^n$. One way of seeing this is to restrict to the case $\mathcal{T} \subseteq [-1,1]^n$. If $\mathcal{T} = [-1,1]^n$ then $\mathtt{Rad}(\mathcal{T}) = 1$, as for any realization of the random signal we can find $t \in \mathcal{T}$ that has its same sign pattern. See **Problem 1.5** in the Problem Sheets for concrete computations of the Rademacher complexity.

2.6 Properties of Rademacher complexity

The Rademacher complexity of a set does not change if all vectors in the set are translated by a constant vector. If all vectors are multiplied by a scalar $c \in \mathbb{R}$, the Rademacher complexity is multiplied by $|c|$.

Proposition 2.6 (Scalar multiplication and translation) *Let $\mathcal{T} \subseteq \mathbb{R}^n$, $v \in \mathbb{R}^n$, $c \in \mathbb{R}$ and define $c\mathcal{T} + v = \{ct + v : t \in \mathcal{T}\}$. Then,*

$$\boxed{\mathtt{Rad}(c\mathcal{T} + v) = |c|\,\mathtt{Rad}(\mathcal{T})}$$

Proof: We have

$$n\,\mathtt{Rad}(\{ct + v : t \in \mathcal{T}\}) = \mathbf{E} \sup_{t\in\mathcal{T}} \sum_{i=1}^n \Omega_i(ct_i + v_i) = \mathbf{E} \sup_{t\in\mathcal{T}} c\sum_{i=1}^n \Omega_i t_i + \mathbf{E}\sum_{i=1}^n \Omega_i v_i = \mathbf{E} \sup_{t\in\mathcal{T}} c\sum_{i=1}^n \Omega_i t_i,$$

where for the last equality we used that $\mathbf{E}\Omega_i = 0$. If $\mathcal{S} \subseteq \mathbb{R}$, then

$$\sup_{x\in\mathcal{S}} c\,x = \begin{cases} |c|\,\sup_{x\in\mathcal{S}} x & \text{if } c \geq 0, \\ c\,\inf_{x\in\mathcal{S}} x = -c\,\sup_{x\in\mathcal{S}}(-x) = |c|\sup_{x\in\mathcal{S}}(-x) & \text{if } c < 0. \end{cases}$$

As each Ω_i has the same distribution than $-\Omega_i$, we have $\mathbf{E}\sup_{t\in\mathcal{T}} c\sum_{i=1}^n \Omega_i t_i = |c|\,\mathbf{E}\sup_{t\in\mathcal{T}} \sum_{i=1}^n \Omega_i t_i$. ■

The Rademacher complexity of a sum of sets is the sum of the Rademacher complexity of the sets.

Proposition 2.7 (Summation) *Let $\mathcal{T}, \mathcal{T}' \subseteq \mathbb{R}^n$ and define $\mathcal{T} + \mathcal{T}' = \{u = t + t' : t \in \mathcal{T}, t' \in \mathcal{T}'\}$. Then,*

$$\boxed{\mathtt{Rad}(\mathcal{T} + \mathcal{T}') = \mathtt{Rad}(\mathcal{T}) + \mathtt{Rad}(\mathcal{T}')}$$

Proof:

$$\mathtt{Rad}(\mathcal{T} + \mathcal{T}') = \mathbf{E} \sup_{t\in\mathcal{T}, t'\in\mathcal{T}'} \sum_{i=1}^n \Omega_i(t_i + t_i') = \mathbf{E} \sup_{t\in\mathcal{T}} \sum_{i=1}^n \Omega_i t_i + \mathbf{E} \sup_{t'\in\mathcal{T}'} \sum_{i=1}^n \Omega_i t_i' = \mathtt{Rad}(\mathcal{T}) + \mathtt{Rad}(\mathcal{T}').$$

■

The Rademacher complexity of a set is the same as the Rademacher complexity of the convex hall of the set. For any $m < \infty$, let Δ_m be the *simplex* in m dimensions, namely, $\Delta_m = \{w = (w_1, \ldots, w_m) \in \mathbb{R}^m : w_1, \ldots, w_m \geq 0, \|w\|_1 = 1\}$.

Proposition 2.8 (Convex Hull) *Let $\mathcal{T} \subseteq \mathbb{R}^n$, and let the convex hull of $\mathcal{T}$ be defined as $\mathrm{conv}(\mathcal{T}) = \{\sum_{j=1}^m w_j t_j : w \in \Delta_m, t_1, \ldots, t_m \in \mathcal{T}, m \in \mathbb{N}\}$. Then,*

$$\boxed{\mathtt{Rad}(\mathrm{conv}(\mathcal{T})) = \mathtt{Rad}(\mathcal{T})}$$

Proof: We have

$$n\,\mathtt{Rad}(\mathrm{conv}(\mathcal{T})) = \mathbf{E}\sup_{m\in\mathbb{N}}\sup_{t_1,\ldots,t_m\in\mathcal{T}}\sup_{w\in\Delta_m}\sum_{i=1}^n\Omega_i\Big(\sum_{j=1}^m w_j t_j\Big)_i = \mathbf{E}\sup_{m\in\mathbb{N}}\sup_{t_1,\ldots,t_m\in\mathcal{T}}\sup_{w\in\Delta_m}\sum_{j=1}^m w_j\sum_{i=1}^n\Omega_i t_{j,i}.$$

Note that for any vector $v=(v_1,\ldots,v_m)\in\mathbb{R}^m$ we have

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

Hence,

$$n\,\mathtt{Rad}(\mathrm{conv}(\mathcal{T})) = \mathbf{E}\sup_{m\in\mathbb{N}}\sup_{t_1,\ldots,t_m\in\mathcal{T}}\max_{j\in 1:m}\sum_{i=1}^n\Omega_i t_{j,i} = \mathbf{E}\sup_{t\in\mathcal{T}}\sum_{i=1}^n\Omega_i t_i = n\,\mathtt{Rad}(\mathcal{T}).$$

■

The Rademacher complexity of a finite set grows at most logarithmically with the set size.

Lemma 2.9 (Finite cardinality, Massart's Lemma) *Let $\mathcal{T}\subseteq\mathbb{R}^n$ and let $\bar{t}:=\frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}t$ be the center of mass. We have*

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

Proof: See **Problem 1.6** in the Problem Sheets. ■

If all vectors in a set are mapped coordinate-wise by γ-Lipschitz functions, the Rademacher complexity is at most multiplied by the Lipschitz constant γ. We recall that a function $f:\mathbb{R}\to\mathbb{R}$ is Lipschitz with parameter γ (or, equivalently, γ-Lipschitz), if $|f(x)-f(y)|\le\gamma|x-y|$ for any $x,y\in\mathbb{R}$.

Let $\mathcal{T}\subseteq\mathbb{R}^n$. Given a function $f:\mathbb{R}\to\mathbb{R}$, we use the notation $f\circ\mathcal{T}$ to denote the subset of $\mathbb{R}^n$ that is obtained by applying the function f to each coordinate of elements in $\mathcal{T}$, namely,

$$f\circ\mathcal{T} := \{(f(t_1),\ldots,f(t_n))\in\mathbb{R}^n : t\in\mathcal{T}\}.$$

Given functions $f_1,\ldots,f_n$ from $\mathbb{R}$ to $\mathbb{R}$, we use the notation $(f_1,\ldots,f_n)\circ\mathcal{T}$ to denote the subset of $\mathbb{R}^n$ that is obtained by applying the functions $f_1,\ldots,f_n$ to the respective coordinates in $\mathcal{T}$, namely,

$$(f_1,\ldots,f_n)\circ\mathcal{T} := \{(f_1(t_1),\ldots,f_n(t_n))\in\mathbb{R}^n : t\in\mathcal{T}\}.$$

Lemma 2.10 (Contraction property, Talagrand's Lemma) *Let $\mathcal{T}\subseteq\mathbb{R}^n$. For each $i\in\{1,\ldots,n\}$, let $f_i:\mathbb{R}\to\mathbb{R}$ be a γ-Lipschitz function. Then,*

$$\boxed{\mathtt{Rad}((f_1,\ldots,f_n)\circ\mathcal{T}) \le \gamma\,\mathtt{Rad}(\mathcal{T})}$$

In particular, if $f:\mathbb{R}\to\mathbb{R}$ is γ-Lipschitz, then

$$\boxed{\mathtt{Rad}(f\circ\mathcal{T}) \le \gamma\,\mathtt{Rad}(\mathcal{T})}$$

Proof: See **Problem 1.7** in the Problem Sheets. ■

2.7 Symmetrization

To see how the Rademacher complexity is related to the problem of finding an upper bound to the quantity $\mathbf{E}\sup_{a\in\mathcal{A}}\{r(a)-R(a)\}$, we use a standard tool in machine learning: symmetrization. We use symmetrization to bound the quantity of interest by the Rademacher complexity of the set in $\mathbb{R}^n$ defined by the composition of the loss function evaluated on the data $S=\{Z_1,\ldots,Z_n\}$ spanned by each possible action in $\mathcal{A}$. We present both a data-dependent upper bound and a data-independent upper bound, by taking the supremum with respect to realization of the data $s=\{z_1,\ldots,z_n\}$.

Let us define the class of functions

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

Given a set of points $\{z_1,\ldots,z_n\}\in\mathcal{Z}^n$, we use the notation $\mathcal{L}\circ\{z_1,\ldots,z_n\}$ to denote the subset of $\mathbb{R}^n$ that is obtained by applying the functions in $\mathcal{L}$ to each element in $\{z_1,\ldots,z_n\}$, namely,

$$\mathcal{L}\circ s:=\{(\ell(a,z_1),\ldots,\ell(a,z_n))\in\mathbb{R}^n:a\in\mathcal{A}\}.$$

Proposition 2.11 *We have*

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

Proof: Let $\{\widetilde{Z}_1,\ldots,\widetilde{Z}_n\}$ be a new sample of independent random variables drawn from the same data distribution, independent of S. Note that by independence and the fact that the $\widetilde{Z}_i$'s are identically distributed we have

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

We have

$$\begin{aligned}
&\mathbf{E}\sup_{a\in\mathcal{A}}\{r(a)-R(a)\}\\
&=\mathbf{E}\sup_{a\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^n\Big(\mathbf{E}[\ell(a,\widetilde{Z}_i)|S]-\ell(a,Z_i)\Big)\\
&=\mathbf{E}\sup_{a\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^n\mathbf{E}[\ell(a,\widetilde{Z}_i)-\ell(a,Z_i)|S]\qquad\text{as }\ell(a,Z_i)=\mathbf{E}[\ell(a,Z_i)|Z_1,\ldots,Z_n]\\
&\le\mathbf{E}\mathbf{E}\Big[\sup_{a\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^n\{\ell(a,\widetilde{Z}_i)-\ell(a,Z_i)\}\Big|S\Big]\\
&=\mathbf{E}\sup_{a\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^n\{\ell(a,\widetilde{Z}_i)-\ell(a,Z_i)\}\quad\text{by the tower property of conditional expectation}\\
&\overset{(a)}{=}\mathbf{E}\sup_{a\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^n\Omega_i\{\ell(a,\widetilde{Z}_i)-\ell(a,Z_i)\}\quad\text{as }(\ell(a,\widetilde{Z}_i)-\ell(a,Z_i))_{i\in[n]}\text{ has same distrib. as }(\Omega_i\{\ell(a,\widetilde{Z}_i)-\ell(a,Z_i))_{i\in[n]}\\
&\le\mathbf{E}\left[\sup_{a\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^n\Omega_i\ell(a,\widetilde{Z}_i)+\sup_{a\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^n(-\Omega_i)\ell(a,Z_i)\right]\quad\text{as }\sup_{a\in\mathcal{A}}(f(a)+g(a))\le\sup_{a\in\mathcal{A}}f(a)+\sup_{a\in\mathcal{A}}g(a)\\
&=2\,\mathbf{E}\sup_{a\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^n\Omega_i\ell(a,Z_i)\quad\text{as }(\Omega_i)_{i\in[n]}\text{ has same distrib. as }(-\Omega_i)_{i\in[n]}\\
&=2\,\mathbf{E}\,\mathtt{Rad}(\mathcal{L}\circ\{Z_1,\ldots,Z_n\}).
\end{aligned}$$

To establish equality (a), we used that for a symmetric random variable X, i.e., $\mathbf{E}\, f(X) = \mathbf{E}\, f(-X)$ for any function f, it holds that if Ω is a Rademacher random variable independent of X, then ΩX has the same distribution as X. To see this formally, note that for any function f we have

$$\begin{aligned}\mathbf{E}\, f(\Omega X) &= \mathbf{E}\,\mathbf{E}[f(\Omega X)|X] = \mathbf{E}[\mathbf{E}[f(\Omega x)]|_{x=X}] = \mathbf{E}\Big[\Big(\frac{1}{2}\mathbf{E}\, f(x) + \frac{1}{2}\mathbf{E}\, f(-x)\Big)\Big|_{x=X}\Big] \\ &= \mathbf{E}\left[\frac{1}{2}\mathbf{E}\, f(X) + \frac{1}{2}\mathbf{E}\, f(-X)\right] = \mathbf{E}\, f(X).\end{aligned}$$

The same argument holds for a collection of independent symmetric random variables $X_1, \ldots, X_n$: $(X_1, \ldots, X_n)$ has the same distribution as $(\Omega_1 X_1, \ldots, \Omega_n X_n)$, namely, for any function f we have

$$\mathbf{E} f(X_1, \ldots, X_n) = \mathbf{E} f(\Omega_1 X_1, \ldots, \Omega_n X_n).$$

Choosing $f(x_1, \ldots, x_n) = \sup_{a\in\mathcal{A}} \frac{1}{n}\sum_{i=1}^n x_i$ and using $X_i = \ell(a, \widetilde{Z}_i) - \ell(a, Z_i)$ concludes the argument. ■

Remark 2.12 (Rademacher complexity of a class of functions) *We have defined the Rademacher complexity of a set, as this is the fundamental object of interest. The (random) quantity*

$$\mathtt{Rad}(\mathcal{L}\circ\{Z_1,\ldots,Z_n\}) = \mathbf{E}\Big[\sup_{a\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^n \Omega_i\ell(a,Z_i)\Big|Z_1,\ldots,Z_n\Big] = \mathbf{E}\Big[\sup_{a\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^n \Omega_i\ell(a,z_i)\Big]\Big|_{z_1=Z_1,\ldots,z_n=Z_n}$$

is typically called the empirical *(or* conditional*) Rademacher complexity of the function class $\mathcal{L}$, and its expectation (a deterministic quantity)*

$$\mathbf{E}\,\mathtt{Rad}(\mathcal{L}\circ\{Z_1,\ldots,Z_n\}) = \mathbf{E}\sup_{a\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^n \Omega_i\ell(a,Z_i)$$

is typically called the Rademacher complexity of the function class $\mathcal{L}$.