Empirical Risk Minimization: Abstract risk bounds and Rademacher averages

Maxim Raginsky

September 24, 2013

In the last lecture, we have left off with a theorem that gave a sufficient condition for the *Empirical Risk Minimization* (ERM) algorithm

$$\widehat{f}_n = \operatorname*{arg\,min}_{f\in\mathscr{F}} \widehat{P}_{Z^n}(\ell_f) \tag{1}$$

$$= \operatorname*{arg\,min}_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) \tag{2}$$

to be PAC for a given learning problem with hypothesis space $\mathscr{F}$ and loss function ℓ. This condition pertained to the behavior of the uniform deviation of empirical means from true means over the *induced class* $\mathscr{L}_{\mathscr{F}} = \{\ell_f : f \in \mathscr{F}\}$. Specifically, we proved that ERM is a PAC algorithm if

$$\lim_{n\to\infty} \sup_{P\in\mathscr{P}} P^n\left(\sup_{f\in\mathscr{F}} \left|\widehat{P}_{Z^n}(\ell_f) - P(\ell_f)\right| \geq \varepsilon\right) = 0, \qquad \forall \varepsilon > 0, \tag{3}$$

where $\mathscr{P}$ is the class of probability distributions generating the training data.

1 An abstract framework for ERM

To study ERM in a general framework, we will adopt a simplified notation often used in the literature. We have a space Z and a class $\mathscr{F}$ of functions $f : \mathsf{Z} \to [0,1]$. Let $\mathscr{P}(\mathsf{Z})$ denote the space of all probability distributions on Z. For each sample size n, the training data are in the form of an n-tuple $Z^n = (Z_1, \ldots, Z_n)$ of Z-valued random variables drawn according to some unknown $P \in \mathscr{P}$. For each P, we can compute the *expected risk* of any $f \in \mathscr{F}$ by

$$P(f) = \mathbb{E}_P f(Z) = \int_{\mathsf{Z}} f(z) P(dz). \tag{4}$$

The *minimum risk* over $\mathscr{F}$ is

$$L_P^*(\mathscr{F}) \triangleq \inf_{f\in\mathscr{F}} P(f). \tag{5}$$

A learning algorithm is a sequence $\mathscr{A} = \{A_n\}_{n\geq 1}$ of mappings $A_n : \mathsf{Z}^n \to \mathscr{F}$, and the objective is to ensure that

$$P(\widehat{f}_n) \approx L_P^*(\mathscr{F}) \qquad \text{eventually with high probability.} \tag{6}$$

The ERM algorithm works by taking

$$\widehat{f}_n = \arg\min_{f\in\mathscr{F}} \widehat{P}_{Z^n}(f) \tag{7}$$

$$= \arg\min_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^n f(Z_i). \tag{8}$$

This way of writing down our problem hides most of the ingredients that were specified in Haussler's framework of model-free learning, so it is important to keep in mind that Z is an input/output pair (X, Y) and the functions $f \in \mathscr{F}$ are really the induced losses for some loss function ℓ and hypothesis class $\mathscr{G}$.

We have already seen that the consistency of ERM hinges on the uniform deviation behavior of empirical means in $\mathscr{F}$. In order to have a clean way of keeping track of all the relevant quantities, let us introduce some additional notation. First of all, we need a way of comparing the behavior of any two probability distributions P and P' on the class $\mathscr{F}$. A convenient way of doing this is through the *$\mathscr{F}$-seminorm*

$$\|P - P'\|_{\mathscr{F}} \triangleq \sup_{f\in\mathscr{F}} |P(f) - P'(f)| \tag{9}$$

$$= \sup_{f\in\mathscr{F}} \left|\mathbb{E}_P f - \mathbb{E}_{P'} f\right| \tag{10}$$

$$= \sup_{f\in\mathscr{F}} \left|\int_{\mathsf{Z}} f(z)P(dz) - \int_{\mathsf{Z}} f(z)P'(dz)\right|. \tag{11}$$

We say that $\|\cdot\|_{\mathscr{F}}$ is a *seminorm* because it has all the properties of a norm (in particular, it satisfies the triangle inequality), but it may happen that $\|P - P'\|_{\mathscr{F}} = 0$ for $P \neq P'$. Next, given a random sample Z^n we define the *uniform deviation*

$$\Delta_n(Z^n) \triangleq \left\|\widehat{P}_{Z^n} - P\right\|_{\mathscr{F}} \equiv \sup_{f\in\mathscr{F}} \left|\widehat{P}_{Z^n}(f) - P(f)\right|. \tag{12}$$

To keep things simple, we do not indicate the underlying distribution P or the function class $\mathscr{F}$ explicitly. We will do this from now on, unless some confusion is possible, in which case we will use appropriate indices. Thus, we will write $L(f)$, $L^*(\mathscr{F})$, etc., and you should always keep in mind that all expectations are computed w.r.t. the (unknown) data-generating distribution $P \in \mathscr{P}(\mathsf{Z})$. In the same spirit, we will denote by $P_n(f)$ the *empirical risk* of f on the sample Z^n:

$$P_n(f) = \widehat{P}_{Z^n}(f) = \frac{1}{n}\sum_{i=1}^n f(Z_i). \tag{13}$$

The following result is key to understanding the role of the uniform deviations $\Delta_n(Z^n)$ in controlling the performance of the ERM algorithm.

Proposition 1. *The ERM algorithm satisfies the following inequalities:*

$$P(\widehat{f}_n) \le L^*(\mathscr{F}) + 2\Delta_n(Z^n) \tag{14}$$

$$P(\widehat{f}_n) \le P_n(\widehat{f}_n) + \Delta_n(Z^n). \tag{15}$$

Proof. We have already proved the two inequalities of the proposition in the last lecture, except now they are written in our new abstract notation. Let us give the proof again in order to get comfortable with the notation. Let f^* be any minimizer of $P(f)$ over $\mathscr{F}$. Then

$$P(\widehat{f}_n) - L^*(\mathscr{F}) = P(\widehat{f}_n) - P(f^*) \tag{16}$$

$$= P(\widehat{f}_n) - P_n(\widehat{f}_n) + P_n(\widehat{f}_n) - P_n(f^*) + P_n(f^*) - P(f^*), \tag{17}$$

where $P_n(\widehat{f}_n) - P_n(f^*) \le 0$ by definition of ERM,

$$P(\widehat{f}_n) - P_n(\widehat{f}_n) \le \sup_{f\in\mathscr{F}} \big[P_n(f) - P(f)\big] \le \|P_n - P\|_{\mathscr{F}} = \Delta_n(Z^n), \tag{18}$$

and the same holds for $P_n(f^*) - P(f^*)$. This proves both (14) and (15). □

The bound (14) says that, if the uniform deviation $\Delta_n(Z^n)$ is small, then the expected risk of the ERM hypothesis will be close to the minimum risk $L^*(\mathscr{F})$; in addition, the bound (15) says that the empirical estimate $P_n(\widehat{f}_n)$ is an accurate estimate of the generalization performance of $\widehat{f}_n$. Both bounds suggest that the success of ERM depends on how small the uniform deviation $\Delta_n(Z^n)$ can be. Thus, we need to develop tools for analyzing the behavior of $\Delta_n(Z^n)$.

2 Bounding the uniform deviation: Rademacher averages

It turns out that the behavior of the uniform deviation $\Delta_n(Z^n)$ is closely connected to how the values of the functions $f \in \mathscr{F}$ on randomly selected n-tuples Z^n correlate with random signs. Intuitively, this can be motivated as follows. In order for ERM to succeed, the function class $\mathscr{F}$ has to be "discriminating:" we should be able to clearly separate all near-minimizers of the empirical risk from functions whose empirical risks (and hence expected risks) are high, but only if the sample is representative of the true data-generating distribution. If the class $\mathscr{F}$ is discriminating not only on the actual sample, but also on its random perturbations, then we cannot expect the empirical risks to truly reflect the generalization ability of the functions in $\mathscr{F}$. As we will soon see, the degree of correlation of the "projections" of $\mathscr{F}$ onto random samples with random signs is captured by the quantities known as the *Rademacher averages*.

First, we need some preparatory results. Let Y be a real-valued random variable. We say that it is *symmetric* if $-Y$ has the same distribution as Y. This is equivalent to saying that

$$\mathbb{P}(Y \ge a) = \mathbb{P}(Y \le -a), \qquad \forall\, a \in \mathbb{R}. \tag{19}$$

A random variable σ taking values -1 or $+1$ with probability $1/2$ is called a *Rademacher random variable*.

Lemma 1. *Let U and U' be two i.i.d. real-valued random variables. Then $Y = U - U'$ is symmetric.*

Proof. Let $F(u) = \mathbb{P}(U \geq u)$. Then

$$\mathbb{P}(Y \geq a) = \mathbb{P}(U - U' \geq a) \tag{20}$$

$$= \mathbb{E}\Big[\mathbb{E}\Big[\mathbf{1}_{\{U-U' \geq a\}} \Big| U'\Big]\Big] \tag{21}$$

$$= \mathbb{E}F(a + U') \tag{22}$$

$$= \mathbb{E}F(a + U), \tag{23}$$

where the last line is because U and U' are i.i.d. An analogous calculation for $\mathbb{P}(Y \leq -a)$ gives the same result. □

Lemma 2. *Let Y be a symmetric random variable, and let σ be a Rademacher random variable independent of Y. Then $W = \sigma Y$ has the same distribution as Y.*

Proof. Direct calculation:

$$\mathbb{P}(W \geq a) = \frac{1}{2}\mathbb{P}(Y \geq a) + \frac{1}{2}\mathbb{P}(Y \leq -a) = \mathbb{P}(Y \geq a), \tag{24}$$

where the second step is due to the symmetry of Y. Since this holds for an arbitrary $a \in \mathbb{R}$, we conclude that W has the same distribution as Y. □

Corollary 1. *Let $Y_1, \ldots, Y_n$ be n independent symmetric random variables, and let $\sigma_1, \ldots, \sigma_n$ be independent Rademacher random variables that are also independent of the Y_i's. Then the sum $Y_1 + \ldots + Y_n$ has the same distribution as $\sigma_1 Y_1 + \ldots + \sigma_n Y_n$.*

Now we are ready to define Rademacher averages.

Definition 1. *The* Rademacher average *of a bounded set $\mathscr{A} \subset \mathbb{R}^n$ is*

$$R_n(\mathscr{A}) \triangleq \mathbb{E} \sup_{a \in \mathscr{A}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i \right|, \tag{25}$$

where the expectation is over n i.i.d. Rademacher random variables $\sigma_1, \ldots, \sigma_n$.

Now consider a class $\mathscr{F}$ of functions $f : \mathsf{Z} \to [0,1]$ from our formulation of the ERM problem. The key result, which we will now prove, is that the uniform deviations $\Delta_n(Z^n)$ are controlled by the Rademacher averages of the *random sets*

$$\mathscr{F}(Z^n) \triangleq \big\{\big(f(Z_1), \ldots, f(Z_n)\big) : f \in \mathscr{F}\big\}. \tag{26}$$

A useful way to think about $\mathscr{F}(Z^n)$ is as a "projection" of $\mathscr{F}$ onto the random sample Z^n.

Theorem 1. *Fix a space Z and let $\mathscr{F}$ be a class of functions $f : \mathsf{Z} \to [0,1]$. Then for any $P \in \mathscr{P}(\mathsf{Z})$*

$$\mathbb{E}\Delta_n(Z^n) \leq 2\mathbb{E}R_n(\mathscr{F}(Z^n)). \tag{27}$$

Proof. The proof uses a clever technique known as "symmetrization," which goes back to the seminal work of Vapnik and Chervonenkis [VC71], but in its modern form is due to Giné and Zinn [GZ84]. The main idea is as follows. Consider a random i.i.d. sample Z^n from P and introduce an independent "ghost" sample $\overline{Z}^n = (\overline{Z}_1, \ldots, \overline{Z}_n)$ from the same P. We will denote expectations w.r.t. $\overline{Z}^n$ by $\overline{\mathbb{E}}$. Let $\overline{P}_n$ denote the empirical distribution of $\overline{Z}^n$. Then for any bounded function $g : \mathsf{Z} \to \mathbb{R}$ we can write

$$P(g) = \overline{\mathbb{E}}\left[\frac{1}{n}\sum_{i=1}^{n} g(\overline{Z}_i)\right] = \overline{\mathbb{E}}\,\overline{P}_n(g). \tag{28}$$

With this, we have

$$\begin{aligned}
\Delta_n(Z^n) &= \|P_n - P\|_{\mathscr{F}} && (29)\\
&= \sup_{f\in\mathscr{F}} \left|P_n(f) - P(f)\right| && (30)\\
&= \sup_{f\in\mathscr{F}} \left|P_n(f) - \overline{\mathbb{E}}\,\overline{P}_n(f)\right| && (31)\\
&\le \sup_{f\in\mathscr{F}} \overline{\mathbb{E}}\left|P_n(f) - \overline{P}_n(f)\right| && (32)\\
&\le \overline{\mathbb{E}} \sup_{f\in\mathscr{F}} \left|P_n(f) - \overline{P}_n(f)\right|, && (33)
\end{aligned}$$

where the first inequality uses convexity of the absolute value function, while the second is because $\sup \overline{\mathbb{E}}[\cdot] \le \overline{\mathbb{E}} \sup[\cdot]$. Now let us take expectations of both sides w.r.t. Z^n to get

$$\mathbb{E}\Delta_n(Z^n) \le \mathbb{E} \sup_{f\in\mathscr{F}} \left|P_n(f) - \overline{P}_n(f)\right|, \tag{34}$$

where now the expectation on the right is w.r.t. both Z^n and $\overline{Z}^n$, which are independent of each other. Let us inspect the difference $P_n(f) - \overline{P}_n(f)$:

$$P_n(f) - \overline{P}_n(f) = \frac{1}{n}\sum_{i=1}^{n}\left[f(Z_i) - f(\overline{Z}_i)\right]. \tag{35}$$

For each i Z_i and $\overline{Z}_i$ are i.i.d., so the differences $f(Z_i) - f(\overline{Z}_i)$ are symmetric by Lemma 1. Introducing n i.i.d. Rademacher random variables $\sigma_1, \ldots, \sigma_n$ independent of Z^n and $\overline{Z}^n$, from Corollary 1 we know that

$$\frac{1}{n}\sum_{i=1}^{n}\left[f(Z_i) - f(\overline{Z}_i)\right] \overset{\mathrm{d}}{=} \frac{1}{n}\sum_{i=1}^{n}\sigma_i\left[f(Z_i) - f(\overline{Z}_i)\right], \tag{36}$$

where $\overset{\mathrm{d}}{=}$ means "equality in distribution." The same holds if we take the supremum of both sides over $f \in \mathscr{F}$. Hence,

$$\begin{aligned}
\mathbb{E}\sup_{f\in\mathscr{F}}\left|P_n(f) - \overline{P}_n(f)\right| &= \mathbb{E}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\left[f(Z_i) - f(\overline{Z}_i)\right]\right| && (37)\\
&= \mathbb{E}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left[f(Z_i) - f(\overline{Z}_i)\right]\right|, && (38)
\end{aligned}$$

where in the last line the expectation is over Z^n, $\overline{Z}^n$, and $\sigma^n = (\sigma_1, \ldots, \sigma_n)$. Now note that

$$\mathbb{E} \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left[f(Z_i) - f(\overline{Z}_i) \right] \right| \leq \mathbb{E} \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i) \right| + \mathbb{E} \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(\overline{Z}_i) \right| \tag{39}$$

$$= 2\mathbb{E} \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i) \right|, \tag{40}$$

where the first line is by the triangle inequality and the second line uses the fact that Z^n has the same distribution as $\overline{Z}^n$. Now, since Z^n and σ^n are independent,

$$\mathbb{E} \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i) \right| = \mathbb{E}_{Z^n} \mathbb{E}_{\sigma^n} \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i) \right| = \mathbb{E} R_n(\mathscr{F}(Z^n)). \tag{41}$$

This completes the proof. □

The above theorem implies the following key result on ERM:

Corollary 2. *For any $P \in \mathscr{P}(\mathsf{Z})$ and any n, the ERM hypothesis $\widehat{f}_n$ satisfies the bound*

$$P(\widehat{f}_n) \leq L^*(\mathscr{F}) + 4\mathbb{E} R_n(\mathscr{F}(Z^n)) + \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}} \tag{42}$$

with probability at least $1-\delta$.

Proof. From Theorem 1 it follows that, for any $t > 0$,

$$\mathbb{P}\Big(\Delta_n(Z^n) \geq 2\mathbb{E} R_n(\mathscr{F}(Z^n)) + t \Big) \leq \mathbb{P}\Big(\Delta_n(Z^n) \geq \mathbb{E}\Delta_n(Z^n) + t \Big). \tag{43}$$

The uniform deviation $\Delta(Z^n)$ has the bounded differences property with $c_1 = \ldots = c_n = 1/n$. Hence, by McDiarmid's inequality

$$\mathbb{P}\Big(\Delta_n(Z^n) \geq \mathbb{E}\Delta_n(Z^n) + t \Big) \leq e^{-2nt^2}.$$

Letting $t = \sqrt{\frac{\log(1/\delta)}{2n}}$, we see that

$$\Delta_n(Z^n) \leq \mathbb{E}\Delta_n(Z^n) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

with probability at least $1-\delta$. Together with (43), this implies that

$$\Delta_n(Z^n) \leq 2\mathbb{E} R_n(\mathscr{F}(Z^n)) + \sqrt{\frac{\log(1/\delta)}{2n}} \tag{44}$$

with probability at least $1-\delta$. Combining this with the first bound of Proposition 1, we conclude that

$$P(\widehat{f}_n) \leq L^*(\mathscr{F}) + 4\mathbb{E} R_n(\mathscr{F}(Z^n)) + \sqrt{\frac{2\log(1/\delta)}{n}} \tag{45}$$

with probability at least $1-\delta$. □

3 Structural results for Rademacher averages

The results developed above highlight the fundamental role played by Rademacher averages in bounding the generalization error of the ERM algorithm. In order to use these bounds, we need to get a better handle on the behavior of Rademacher averages.

Lemma 3 (Basic properties of Rademacher averages)**.** *Let $\mathscr{A}$ and $\mathscr{B}$ be bounded subsets of $\mathbb{R}^n$, and let $c \in \mathbb{R}$ be a constant. Then*

$$R_n(\mathscr{A} \cup \mathscr{B}) \leq R_n(\mathscr{A}) + R_n(\mathscr{B}) \tag{46}$$

$$R_n(c\mathscr{A}) = |c| R_n(\mathscr{A}) \tag{47}$$

$$R_n(\mathscr{A} + \mathscr{B}) \leq R_n(\mathscr{A}) + R_n(\mathscr{B}), \tag{48}$$

where $c\mathscr{A} \triangleq \{ca : a \in \mathscr{A}\}$ and $\mathscr{A} + \mathscr{B} \triangleq \{a + b : a \in \mathscr{A}, b \in \mathscr{B}\}$. Moreover, let

$$\operatorname{conv}\mathscr{A} \triangleq \left\{ \sum_{m=1}^{N} c_m a_m : N \in \mathbb{N}; a_m \in \mathscr{A}; c_m \geq 0, \forall m; \sum_{m=1}^{N} c_m = 1 \right\} \tag{49}$$

be the convex hull of $\mathscr{A}$ and

$$\operatorname{absconv}\mathscr{A} \triangleq \left\{ \sum_{m=1}^{N} c_m a_m : N \in \mathbb{N}; a_m \in \mathscr{A}; \sum_{m=1}^{n} |c_m| \leq 1 \right\} \tag{50}$$

be the absolute convex hull *of $\mathscr{A}$. Then*

$$R_n(\mathscr{A}) = R_n(\operatorname{conv}\mathscr{A}) = R_n(\operatorname{absconv}\mathscr{A}). \tag{51}$$

Proof. The proof is by direct calculation. First of all, by double-counting,

$$R_n(\mathscr{A} \cup \mathscr{B}) = \mathbb{E} \sup_{v \in \mathscr{A} \cup \mathscr{B}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i v_i \right| \tag{52}$$

$$\leq \mathbb{E} \sup_{a \in \mathscr{A}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i \right| + \mathbb{E} \sup_{b \in \mathscr{B}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i b_i \right| \tag{53}$$

$$= R_n(\mathscr{A}) + R_n(\mathscr{B}). \tag{54}$$

The case of $c\mathscr{A}$ is obvious. For $\mathscr{A} + \mathscr{B}$,

$$R_n(\mathscr{A} + \mathscr{B}) = \mathbb{E} \sup_{v \in \mathscr{A} + \mathscr{B}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i v_i \right| \tag{55}$$

$$= \mathbb{E} \sup_{a \in \mathscr{A}, b \in \mathscr{B}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i (a_i + b_i) \right| \tag{56}$$

$$\leq \mathbb{E} \sup_{a \in \mathscr{A}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i \right| + \mathbb{E} \sup_{b \in \mathscr{B}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i b_i \right| \tag{57}$$

$$= R_n(\mathscr{A}) + R_n(\mathscr{B}), \tag{58}$$

where the third step uses the triangle inequality.

Finally, consider the absolute convex hull of $\mathscr{A}$. Since $\mathscr{A} \subset \operatorname{absconv}\mathscr{A}$, $R_n(\mathscr{A}) \le R_n(\operatorname{absconv}\mathscr{A})$. On the other hand, fix some $N \in \mathbb{N}$ and N real numbers $c_1, \ldots, c_N$ such that $\sum_{m=1}^N |c_m| = 1$, and consider the set

$$c_1\mathscr{A} + \ldots + c_N\mathscr{A} \equiv \{c_1 a_1 + \ldots + c_N a_N : a_1, \ldots, a_N \in \mathscr{A}\}. \tag{59}$$

Then

$$R_n(c_1\mathscr{A} + \ldots + c_N\mathscr{A}) \le \sum_{i=1}^N |c_i| R_n(\mathscr{A}) \le R_n(\mathscr{A}). \tag{60}$$

Since $\operatorname{absconv}\mathscr{A}$ is the union of all sets of the form (59) for all choices of N and $\{c_m\}_{m=1}^N$, we see that $R_n(\operatorname{absconv}\mathscr{A}) \le R_n(\mathscr{A})$. Therefore, $R_n(\mathscr{A}) = R_n(\operatorname{absconv}\mathscr{A})$. Since $\mathscr{A} \subset \operatorname{conv}\mathscr{A} \subset \operatorname{absconv}\mathscr{A}$, the same equality holds for the convex hull of $\mathscr{A}$. □

The properties listed in the lemma show what happens to Rademacher averages when we form combinations of sets. This will be useful to us later, when we talk about hypothesis classes made up of simpler classes by means of operations like set-theoretic unions, intersections, complements or differences, logical connectives, or convex and linear combinations.

The next result, often referred to as the Finite Class Lemma, is very important:

Lemma 4 (Finite class lemma)**.** *If $\mathscr{A} = \{a^{(1)}, \ldots, a^{(N)}\} \subset \mathbb{R}^n$ is a finite set with $\|a^{(j)}\| \le L$ for all $j = 1, \ldots, N$ and $N \ge 2$, then*

$$R_n(\mathscr{A}) \le \frac{2L\sqrt{\log N}}{n}. \tag{61}$$

Proof. Let σ^n be n i.i.d. Rademacher variables, and for every $j \in 1, \ldots, N$ let

$$Y_j \triangleq \frac{1}{n}\sum_{i=1}^n \sigma_i a_i^{(j)}. \tag{62}$$

Then for any $s > 0$

$$\mathbb{E}e^{sY_j} = \mathbb{E}\exp\left(\frac{s}{n}\sum_{i=1}^n \sigma_i a_i^{(j)}\right) = \prod_{i=1}^n \mathbb{E}e^{s\sigma_i a_i^{(j)}/n}, \tag{63}$$

where the second step uses the fact that σ^n are i.i.d. For each i, the random variable $\sigma_i a_i^{(j)}$ has zero mean and is bounded between $-a_i^{(j)}$ and $a_i^{(j)}$, so by the Hoeffding bound we have

$$\mathbb{E}e^{s\sigma_i a_i^{(j)}/n} \le \exp\left(\frac{s^2\left|a_i^{(j)}\right|^2}{2n^2}\right). \tag{64}$$

Therefore,

$$\mathbb{E}e^{sY_j} \le \prod_{i=1}^{n} \exp\left(\frac{s^2\left|a_i^{(j)}\right|^2}{2n^2}\right) = \exp\left(\frac{s^2}{2n^2}\sum_{i=1}^{n}\left|a_i^{(j)}\right|^2\right) = \exp\left(\frac{s^2\|a^{(j)}\|^2}{2n^2}\right) \le \exp\left(\frac{s^2L^2}{2n^2}\right). \tag{65}$$

Repeating the same argument for each $-Y_j$, we see that

$$\mathbb{E}e^{-sY_j} \le \exp\left(\frac{s^2L^2}{2n^2}\right). \tag{66}$$

Now we recall the following statement, which was given as a homework problem in Spring 2011 at Duke University[1]: Let $U_1, \ldots, U_K$ be K random variables (not necessarily independent) that are *subgaussian* with parameter $v > 0$, i.e.,

$$\mathbb{E}[e^{sU_k}] \le e^{s^2v^2/2}, \qquad \forall s > 0. \tag{67}$$

Then

$$\mathbb{E}\left[\max_{1\le k\le K} U_k\right] \le v\sqrt{2\log K}. \tag{68}$$

Consider now the $2N$ random variables $Y_1, -Y_1, \ldots, Y_N, -Y_N$. According to (65) and (66), they are subgaussian with parameter $v = L/n$. Hence,

$$\mathbb{E}\left[\max_{1\le j\le N}|Y_j|\right] = \mathbb{E}\left[\max\left(Y_1, -Y_1, \ldots, Y_N, -Y_N\right)\right] \le \frac{L\sqrt{2\log(2N)}}{n} \le \frac{2L\sqrt{\log N}}{n}, \tag{69}$$

where the last step uses the fact that, since $N \ge 2$, $2N \le N^2$. Finally,

$$R_n(\mathscr{A}) = \mathbb{E}\left[\max_{1\le j\le N}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i a_i^{(j)}\right|\right] = \mathbb{E}\left[\max_{1\le j\le N}|Y_j|\right] \le \frac{2L\sqrt{\log N}}{n}, \tag{70}$$

which is what we wanted to prove. □

We will start exploring the implications of the Finite Class Lemma more fully in the next lecture, but we can give a brief preview here. Consider a learning problem of the type described in Section 1 in the special case when $\mathscr{F}$ consists of *binary-valued* functions on Z, i.e., $\mathscr{F} = \left\{f : \mathsf{Z} \to \{0,1\}\right\}$. From Theorem 1, we know that

$$\mathbb{E}\Delta_n(Z^n) \le 2\mathbb{E}R_n(\mathscr{F}(Z^n)), \tag{71}$$

where

$$\mathscr{F}(Z^n) \triangleq \left\{\left(f(Z_1), \ldots, f(Z_n)\right) : f \in \mathscr{F}\right\}. \tag{72}$$

[1] See Problem 2 in `http://maxim.ece.illinois.edu/teaching/spring11/homework/homework_1.pdf`

Note that because each f can take values 0 or 1, $\mathscr{F}(Z^n) \subseteq \{0,1\}^n$. Moreover, since for any $Z^n \in \mathsf{Z}^n$ and any $f \in \mathscr{F}$ we have

$$\sqrt{\sum_{i=1}^{n} |f(Z_i)|^2} \leq \sqrt{n}, \tag{73}$$

the set $\mathscr{F}(Z^n)$ for a *fixed* Z^n satisfies the conditions of the Finite Class Lemma with $N = |\mathscr{F}(Z^n)| \leq 2^n$ and $L = \sqrt{n}$. Hence,

$$R_n(\mathscr{F}(Z^n)) \leq 2\sqrt{\frac{\log|\mathscr{F}(Z^n)|}{n}}. \tag{74}$$

In general, since $\log|\mathscr{F}(Z^n)| \leq n$, the bound just says that $R_n(\mathscr{F}(Z^n)) \leq 2$, which is not that useful. However, as we will see in the next few lectures, for a broad range of binary function classes $\mathscr{F}$ it will not be possible to pick out every single element in $\{0,1\}^n$ by taking the random "slices" $\mathscr{F}(Z^n)$, provided n is sufficiently large. To make these notions precise, let us define the quantity

$$\mathbb{S}_n(\mathscr{F}) \triangleq \sup_{z^n \in \mathsf{Z}^n} |\mathscr{F}(z^n)|, \tag{75}$$

which is called the *nth shatter coefficient* of $\mathscr{F}$. Then we have the bound

$$R_n(\mathscr{F}(Z^n)) \leq 2\sqrt{\frac{\log\mathbb{S}_n(\mathscr{F})}{n}}. \tag{76}$$

Next, let

$$V(\mathscr{F}) \triangleq \max\left\{n \in \mathbb{N} : \mathbb{S}_n(\mathscr{F}) = 2^n\right\}. \tag{77}$$

This number is the famous *Vapnik–Chervonenkis* (or *VC*) *dimension* of $\mathscr{F}$, which has originated in their work [VC71]. It is clear that if $\mathbb{S}_n(\mathscr{F}) < 2^n$ for some n, then $\mathbb{S}_m(\mathscr{F}) < 2^m$ for all $m > n$. Hence, $V(\mathscr{F})$ is always well-defined (though it may be infinite). When it is finite, we say that $\mathscr{F}$ is a *VC class*. What this means is that, for n large enough, a certain structure emerges in the sets $\mathscr{F}(z^n)$, which prevents us from being able to form any combination of binary labels by sweeping through the entire $\mathscr{F}$. A fundamental result, which was independently derived by Sauer [Sau72] and Shelah [She72] in different contexts (combinatorics and mathematical logic respectively) and also appeared in a weaker form in the original work of Vapnik and Chervonenkis [VC71], says the following:

Lemma 5 (Sauer–Shelah)**.** *If $\mathscr{F}$ is a VC class, i.e., $V(\mathscr{F}) < \infty$, then*

$$\mathbb{S}_n(\mathscr{F}) \leq \sum_{i=1}^{V(\mathscr{F})} \binom{n}{i} \leq (n+1)^{V(\mathscr{F})}. \tag{78}$$

Thus, we arrive at the following important result, which we will revisit in the next lecture:

Theorem 2. *If $\mathscr{F}$ is a VC class of binary functions, then*

$$\mathbb{E}R_n(\mathscr{F}(Z^n)) \le 2\sqrt{\frac{V(\mathscr{F})\log(n+1)}{n}}. \tag{79}$$

Consequently, for a VC class $\mathscr{F}$, the risk of ERM computed on an i.i.d. sample of size n from an arbitrary distribution $P \in \mathscr{P}(\mathsf{Z})$ is bounded by

$$P(\widehat{f}_n) \le L^*(\mathscr{F}) + 8\sqrt{\frac{V(\mathscr{F})\log(n+1)}{n}} + \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}} \tag{80}$$

with probability at least $1-\delta$. In fact, using a much more refined technique called *chaining* originating in the work of Dudley [Dud78], it is possible to remove the logarithm in (79) to obtain the bound

$$\mathbb{E}R_n(\mathscr{F}(Z^n)) \le C\sqrt{\frac{V(\mathscr{F})}{n}}, \tag{81}$$

where $C > 0$ is some universal constant independent of n and $\mathscr{F}$. We will not cover chaining in this class, but we will use the above formula.

References

[Dud78] R. M. Dudley. Central limit theorems for empirical measures. *Annals of Probability*, 6:899–929, 1978.

[GZ84] E. Giné and J. Zinn. Some limit theorems for empirical processes. *The Annals of Probability*, 12:929–989, 1984.

[Sau72] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13:145–147, 1972.

[She72] S. Shelah. A combinatorial problem: stability and order for models and theories in infinity languages. *Pacific Journal of Mathematics*, 41:247–261, 1972.

[VC71] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16:264–280, 1971.