IEMS 402: Statistical Learning **Winter 2024-2025**

Lecture 9: Rademacher complexity

Lecturer: Yiping Lu *Scribes: Yuanxin Liu*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

9.1 Definitions

Given a space Z and a fixed distribution D_Z, let $S = z_1, z_2, \ldots, z_m$ be a set of examples drawn i.i.d. from D_Z. Furthermore, let F be a class of functions $f : Z \to \mathbb{R}$.

Definition 9.1 (Empirical Rademacher Complexity) *The empirical Rademacher complexity of $\mathcal{F}$ is defined as*

$$\hat{R}_m(\mathcal{F}) = \mathbb{E}\left[\sup_{f \in F} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i)\right],$$

where $\sigma_1, \sigma_2, \ldots, \sigma_m$ are independent random variables uniformly chosen from $\{-1, 1\}$, known as Rademacher variables.

In this definition, it is important to note the position of the expectation and supremum. If the supremum is taken outside the expectation, the result is 0 since the expectation of Rademacher variables is 0.

Definition 9.2 (Rademacher Complexity) *The Rademacher complexity of $\mathcal{F}$ is defined as*

$$R_m(\mathcal{F}) = \mathbb{E}_D[\hat{R}_m(\mathcal{F})].$$

Intuitively, the supremum in the definition measures, for a given set S and a Rademacher vector σ, the maximum correlation between $f(z_i)$ and σ_i over all $f \in \mathcal{F}$. Taking the expectation over σ, we can say that the empirical Rademacher complexity of $\mathcal{F}$ quantifies the ability of functions in $\mathcal{F}$ (applied to a fixed set S) to fit random noise. The Rademacher complexity of $\mathcal{F}$ then measures the expected noise-fitting ability of $\mathcal{F}$ over all possible data sets $S = (z_1, z_2, \ldots, z_m)$ that could be drawn according to the distribution D_Z. Note that Rademacher complexity can be defined more generally for sets $A \subset \mathbb{R}^m$ by taking the supremum over A (instead of $\mathcal{F}$) and replacing each $f(z_i)$ with a_i. Taking $A = F(S) = \{f(z) \mid f \in \mathcal{F}, z \in S\}$ recovers the definition above. It will sometimes be convenient to use this more general definition.

9.2 Generalization Bound via Rademacher Complexity

Theorem 9.3 *Fix a distribution D_Z and a parameter $\delta \in (0, 1)$. If $\mathcal{F} \subset \{f : Z \to [a, a+1]\}$ and $S = \{z_1, \ldots, z_n\}$ is drawn i.i.d. from D_Z, then with probability at least $1 - \delta$ over the draw of S, for every function $f \in \mathcal{F}$,*

$$\mathbb{E}_D[f(z)] \leqslant \hat{E}_S[f(z)] + 2R_m(\mathcal{F}) + \sqrt{\frac{\ln(\frac{1}{\delta})}{m}} \tag{1}$$

where $\hat{E}_S[f(z)] := \frac{1}{m}\sum_{i=1}^m f(z_i)$*, and* $R_m(\mathcal{F})$ *is the Rademacher complexity of* $\mathcal{F}$*.*

In addition, with probability at least $1-\delta$*, for every function* $f \in \mathcal{F}$*,*

$$\mathbb{E}_D[f(z)] \leqslant \hat{E}_S[f(z)] + 2\hat{R}_m(\mathcal{F}) + 3\sqrt{\frac{\ln(\frac{2}{\delta})}{m}} \tag{2}$$

where $\hat{R}_m(\mathcal{F})$ *is the empirical Rademacher complexity computed from the sample* S*.*

In what follows we prove two key theorems.

9.2.1 Symmetrization

Lemma 9.4 (Symmetrization) *Let* P *be a probability distribution over a domain* Z*. The Rademacher complexity of the function class* $\mathcal{F}$ *with respect to* P*, for an i.i.d. sample* $S = \{z_1, \ldots, z_m\}$ *of size* m*, is given by* $R_m(\mathcal{F})$*. Then,*

$$\mathbb{E}_S \sup_{f\in\mathcal{F}} \left(\mathbb{E}_{z\sim P}[f(z)] - \frac{1}{m}\sum_{i=1}^m f(z_i)\right) \leqslant 2\,R_m(\mathcal{F}).$$

Proof: We start by writing the quantity of interest:

$$\Phi(S) := \sup_{f\in\mathcal{F}} \left(\mathbb{E}[f(z)] - \frac{1}{m}\sum_{i=1}^m f(z_i)\right).$$

Let $S' = \{z'_1, \ldots, z'_m\}$ be an independent copy of S, i.e., the z'_i are also drawn i.i.d. from P. Note that

$$\mathbb{E}_{z\sim P}[f(z)] = \mathbb{E}_{S'}\left[\frac{1}{m}\sum_{i=1}^m f(z'_i)\right].$$

Thus,

$$\Phi(S) = \sup_{f\in\mathcal{F}} \left(\mathbb{E}_{S'}\frac{1}{m}\sum_{i=1}^m f(z'_i) - \frac{1}{m}\sum_{i=1}^m f(z_i)\right).$$

By exchanging the order of the supremum and expectation (via Jensen's inequality) we have

$$\Phi(S) \leqslant \mathbb{E}_{S'}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m (f(z'_i) - f(z_i))\right].$$

Now, by the linearity of expectation and using the fact that the two samples S and S' are identically distributed, we introduce Rademacher variables $\sigma_1, \ldots, \sigma_m$ and note that for any fixed pair (z_i, z'_i) the pair $(f(z'_i) - f(z_i))$ is symmetric in distribution. Thus, we can write:

$$\mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m (f(z'_i) - f(z_i))\right] = \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i\,(f(z'_i) - f(z_i))\right].$$

Using the triangle inequality and the fact that the distribution of (z_i) and (z'_i) are the same, we obtain:

$$\mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m (f(z'_i) - f(z_i))\right] \leqslant \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i f(z'_i)\right] + \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i f(z_i)\right].$$

Since both terms are equal by symmetry, we conclude that

$$\mathbb{E}_S[\Phi(S)] \leqslant 2\,\mathbb{E}_S\,\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^m \sigma_i f(z_i)\right] = 2\,R_m(\mathcal{F}).$$

This completes the proof. ∎

9.2.2 Concentration for Rademacher Complexities and Estimation Error

Lemma 9.5 *Let $\mathcal{F}$ be a set of functions such that for any $f \in \mathcal{F}$ and for any two points x, y in the domain of f, $|f(x) - f(y)| \leqslant c$, for some constant c. Let $R_m(\mathcal{F})$ and $\hat{R}_m(\mathcal{F}_S)$ be the Rademacher complexity and the empirical Rademacher complexity of $\mathcal{F}$ with respect to an i.i.d. sample $S = \{z_1, \ldots, z_m\}$ drawn from P. Then:*

1. *For any $\epsilon > 0$,*

$$P(\hat{R}_m(\mathcal{F}_S) - R_m(\mathcal{F}) \geqslant \epsilon) \leqslant 2\exp\left(-\frac{2m^2\epsilon^2}{c^2}\right),$$

and

$$P(R_m(\mathcal{F}) - \hat{R}_m(\mathcal{F}_S) \geqslant \epsilon) \leqslant 2\exp\left(-\frac{2m^2\epsilon^2}{c^2}\right).$$

2. *For all $f \in \mathcal{F}$ and for any $\epsilon > 0$,*

$$P(E[f(z)] - \hat{E}_S[f(z)] \geqslant 2\,\hat{R}_m(\mathcal{F}_S) + \epsilon) \leqslant 2\exp\left(-\frac{2m^2\epsilon^2}{c^2}\right).$$

Proof: 1. Concentration of the Empirical Rademacher Complexity:
The empirical Rademacher complexity is defined as

$$\hat{R}_m(\mathcal{F}_S) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^m \sigma_i f(z_i)\right].$$

Because each function f is c-Lipschitz (with respect to its output) and each $f(z_i)$ is in an interval of length at most c, a change in a single sample z_i can change the value of $\hat{R}_m(\mathcal{F}_S)$ by at most $\frac{c}{m}$. Hence, McDiarmid's inequality implies that for any $\epsilon > 0$,

$$P(\hat{R}_m(\mathcal{F}_S) - \mathbb{E}_S[\hat{R}_m(\mathcal{F}_S)] \geqslant \epsilon) \leqslant \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^m \left(\frac{c}{m}\right)^2}\right) = \exp\left(-\frac{2m^2\epsilon^2}{mc^2}\right) = \exp\left(-\frac{2m\epsilon^2}{c^2}\right).$$

A similar bound holds for the lower tail. Since by definition,

$$\mathbb{E}_S[\hat{R}_m(\mathcal{F}_S)] = R_m(\mathcal{F}),$$

we obtain the stated bounds with an extra factor 2 (by a standard symmetrization of the two tails):

$$P(|\hat{R}_m(\mathcal{F}_S) - R_m(\mathcal{F})| \geqslant \epsilon) \leqslant 2\exp\left(-\frac{2m^2\epsilon^2}{c^2}\right).$$

2. Concentration for the Estimation Error:

We wish to bound the deviation

$$\sup_{f\in\mathcal{F}}\left(E[f(z)]-\hat{E}_S[f(z)]\right).$$

From Lemma 9.4 (the symmetrization result) we have

$$\mathbb{E}_S \sup_{f\in\mathcal{F}}(E[f(z)]-\hat{E}_S[f(z)]) \leqslant 2\,R_m(\mathcal{F}).$$

Now, using the fact that each $f(z)$ is bounded in an interval of length at most c, a change in one sample z_i changes

$$\frac{1}{m}\sum_{i=1}^{m} f(z_i)$$

by at most $\frac{c}{m}$. Hence, McDiarmid's inequality can be applied directly to the function

$$\phi(S)=\sup_{f\in\mathcal{F}}\left(E[f(z)]-\hat{E}_S[f(z)]\right).$$

Thus, for any $\epsilon>0$,

$$P(\sup_{f\in\mathcal{F}}(E[f(z)]-\hat{E}_S[f(z)]) \geqslant \mathbb{E}_S[\phi(S)]+\epsilon) \leqslant \exp\left(-\frac{2m^2\epsilon^2}{c^2}\right).$$

Using the concentration result from part (1) to relate $R_m(\mathcal{F})$ with the empirical counterpart $\hat{R}_m(\mathcal{F}_S)$ (i.e., with high probability,

$$R_m(\mathcal{F}) \leqslant \hat{R}_m(\mathcal{F}_S)+\epsilon_1,$$

with $\epsilon_1=\sqrt{\frac{\ln(2/\delta)}{2m^2/c^2}}$), one can absorb the additional deviation into the bound. In particular, by choosing parameters appropriately (and possibly relaxing the constants), we obtain that for any $\epsilon>0$,

$$P(\sup_{f\in\mathcal{F}}(E[f(z)]-\hat{E}_S[f(z)]) \geqslant 2\,\hat{R}_m(\mathcal{F}_S)+\epsilon) \leqslant 2\exp\left(-\frac{2m^2\epsilon^2}{c^2}\right).$$

This completes the proof of Lemma 9.5. ■

9.2.3 Derivation of the Generalization Bounds (1) and (2)

Using Lemma 9.4 we have for any $f\in\mathcal{F}$,

$$E[f(z)] \leqslant \hat{E}_S[f(z)]+\sup_{f\in\mathcal{F}}(E[f(z)]-\hat{E}_S[f(z)]).$$

Taking expectation over the sample and then applying Lemma 9.4 yields

$$\mathbb{E}_S\left[E[f(z)]-\hat{E}_S[f(z)]\right] \leqslant 2\,R_m(\mathcal{F}).$$

By applying McDiarmid's inequality (as in the proofs above) to control the deviation from the expectation, we conclude that with probability at least $1-\delta$,

$$E[f(z)] \leqslant \hat{E}_S[f(z)]+2\,R_m(\mathcal{F})+\sqrt{\frac{\ln(1/\delta)}{m}}.$$

This is the generalization bound (1).

Next, using the concentration result from Lemma 9.5 that relates the true and the empirical Rademacher complexities, namely that with high probability

$$R_m(\mathcal{F}) \leqslant \hat{R}_m(\mathcal{F}_S) + \sqrt{\frac{\ln(2/\delta)}{m}},$$

substitute the above into the bound (1) to get

$$E[f(z)] \leqslant \hat{E}_S[f(z)] + 2\hat{R}_m(\mathcal{F}_S) + 2\sqrt{\frac{\ln(2/\delta)}{m}} + \sqrt{\frac{\ln(1/\delta)}{m}}.$$

By slightly relaxing the constants (noting that $\sqrt{\frac{\ln(1/\delta)}{m}} \leqslant \sqrt{\frac{\ln(2/\delta)}{m}}$ for $\delta < 1$), we obtain

$$E[f(z)] \leqslant \hat{E}_S[f(z)] + 2\hat{R}_m(\mathcal{F}_S) + 3\sqrt{\frac{\ln(2/\delta)}{m}},$$

which is the generalization bound (2).

9.3 Bound Rademacher Complexity by Covering Number

Theorem 9.6 (Massart's Lemma) *Assume that $\mathcal{F}$ is finite. Let $S = \{z_1, z_2, \ldots, z_m\}$ be a random i.i.d. sample, and let $B = \max_{f\in\mathcal{F}} \left(\frac{1}{m}\sum_{i=1}^m f^2(z_i)\right)^{\frac{1}{2}}$. Then, the empirical Rademacher complexity satisfies*

$$\hat{R}_m(\mathcal{F}_S) \leqslant B\sqrt{\frac{2\ln|\mathcal{F}|}{m}}.$$

Proof: For any $s > 0$, we start with

$$\exp(s\, m\, R_m(\mathcal{F}_S)) = \exp\left(s\,\mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^m \varepsilon_i f(z_i)\right]\right),$$

where $\{\varepsilon_i\}_{i=1}^m$ are independent Rademacher random variables. By Jensen's inequality,

$$\exp\left(s\,\mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^m \varepsilon_i f(z_i)\right]\right) \leqslant \mathbb{E}\left[\sup_{f\in\mathcal{F}} \exp\left(s\sum_{i=1}^m \varepsilon_i f(z_i)\right)\right].$$

Since the supremum is over a finite set, we can bound the expectation by summing over $\mathcal{F}$:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \exp\left(s\sum_{i=1}^m \varepsilon_i f(z_i)\right)\right] \leqslant \sum_{f\in\mathcal{F}} \prod_{i=1}^m \mathbb{E}\left[\exp\left(s\varepsilon_i f(z_i)\right)\right].$$

By Hoeffding's lemma, since $\mathbb{E}[\varepsilon_i] = 0$, we have

$$\mathbb{E}\left[\exp\left(s\varepsilon_i f(z_i)\right)\right] \leqslant \exp\left(\frac{s^2 f^2(z_i)}{2}\right).$$

Thus,

$$\prod_{i=1}^m \mathbb{E}\left[\exp\left(s\varepsilon_i f(z_i)\right)\right] \leqslant \exp\left(\frac{s^2}{2}\sum_{i=1}^m f^2(z_i)\right).$$

Taking the maximum over $\mathcal{F}$, we obtain

$$\exp(s\,m\,R_m(\mathcal{F}_S)) \leqslant |\mathcal{F}| \exp\left(\frac{s^2\,m\,B^2}{2}\right).$$

Taking logarithms and dividing by m yields

$$R_m(\mathcal{F}_S) \leqslant \frac{1}{s\,m} \ln|\mathcal{F}| + \frac{sB^2}{2}.$$

Optimizing over s, choose

$$s = \sqrt{\frac{2\ln|\mathcal{F}|}{mB^2}},$$

which, when substituted back, gives

$$R_m(\mathcal{F}_S) \leqslant B\sqrt{\frac{2\ln|\mathcal{F}|}{m}}.$$

■

Theorem 9.7 (Covering Number Bound) *Let $\mathcal{F}$ be a class of real-valued functions, let $S = \{z_1, z_2, \ldots, z_m\}$ be a random i.i.d. sample, and let $C(\mathcal{F}, \|\cdot\|_{1,S})$ denote the size of a minimal cover of $\mathcal{F}$ with respect to the $\ell_1(S)$-norm (i.e., the covering number). Assuming that*

$$\sup_{f\in\mathcal{F}} \left(\frac{1}{m}\sum_{i=1}^{m} f^2(z_i)\right)^{1/2} \leqslant c,$$

we have

$$R_m(\mathcal{F}_S) \leqslant \inf_{\epsilon>0} \left\{\epsilon + \frac{\sqrt{2}c}{\sqrt{m}}\sqrt{\ln C(\mathcal{F}, \|\cdot\|_{1,S})}\right\}.$$

Proof: Fix any $\epsilon > 0$. Let F be a minimal ϵ–cover of $\mathcal{F}$ with respect to the norm $\|\cdot\|_{1,S}$, i.e., for any $f \in \mathcal{F}$ there exists $f' \in F$ such that

$$\frac{1}{m}\sum_{i=1}^{m}|f(z_i) - f'(z_i)| < \epsilon.$$

Note that by definition, F is an ϵ–cover of $\mathcal{F}$. Then, writing the Rademacher complexity of $\mathcal{F}_S$ as

$$R_m(\mathcal{F}_S) = \frac{1}{m}\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i f(z_i),$$

we decompose each $f \in \mathcal{F}$ as

$$f(z_i) = (f(z_i) - f'(z_i)) + f'(z_i)$$

for some $f' \in F$. Hence,

$$R_m(\mathcal{F}_S) = \frac{1}{m}\mathbb{E}\sup_{f\in\mathcal{F}}\left\{\sum_{i=1}^{m}\sigma_i(f(z_i) - f'(z_i)) + \sum_{i=1}^{m}\sigma_i f'(z_i)\right\}. \tag{9.1}$$

For clarity, denote the two terms by

$$A = \frac{1}{m}\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i(f(z_i) - f'(z_i))$$

and

$$B = \frac{1}{m}\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i f'(z_i).$$

Note that the supremum in (9.1) is taken over all $f \in \mathcal{F}$, and for each f the corresponding f' depends on f. Thus, we cannot exchange the supremum and the summation in B.

We now bound the terms A and B separately.

Term A. By the covering property, for any $f \in \mathcal{F}$ we have

$$\frac{1}{m}\sum_{i=1}^{m}|f(z_i) - f'(z_i)| < \epsilon.$$

Since the Rademacher variables σ_i satisfy $|\sigma_i|= 1$, it follows that

$$\left|\sum_{i=1}^{m}\sigma_i(f(z_i) - f'(z_i))\right| \leqslant \sum_{i=1}^{m}|f(z_i) - f'(z_i)| < m\epsilon.$$

Thus,

$$A \leqslant \frac{1}{m}\cdot m\epsilon = \epsilon.$$

Term B. Since F is a finite cover of $\mathcal{F}$ with covering number

$$C(\mathcal{F}, \|\cdot\|_{1,S}),$$

standard bounds on the Rademacher complexity (via Massart's lemma or similar arguments) yield

$$B = \frac{1}{m}\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i f'(z_i) \leqslant \frac{\sqrt{2}\,c}{\sqrt{m}}\sqrt{\ln C(\mathcal{F}, \|\cdot\|_{1,S})},$$

where c is an absolute constant and we have, as usual, replaced $R(\mathcal{F}, S)$ by $R(\mathcal{F}_S)$.

Combining the bounds for A and B, we obtain

$$R_m(\mathcal{F}_S) \leqslant \epsilon + \frac{\sqrt{2}\,c}{\sqrt{m}}\sqrt{\ln C(\mathcal{F}, \|\cdot\|_{1,S})}.$$

Since the above inequality holds for any $\epsilon > 0$, we conclude that

$$R_m(\mathcal{F}_S) \leqslant \inf_{\epsilon>0}\left\{\epsilon + \frac{\sqrt{2}\,c}{\sqrt{m}}\sqrt{\ln C(\mathcal{F}, \|\cdot\|_{1,S})}\right\}.$$

■

Theorem 9.8 (Dudley's Entropy Integral Bound) *Let $\mathcal{F}$ be a class of real-valued functions, let $S = \{z_1, z_2, \ldots, z_m\}$ be a random i.i.d. sample, and let $C(\mathcal{F}, \epsilon, \|\cdot\|_{2,\mathcal{S}})$ denote the size of a minimal cover of $\mathcal{F}$ with respect to the $\|\cdot\|_{2,\mathcal{S}}$. Assuming that $\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m} f^2(z_i)\right)^{\frac{1}{2}} \leqslant c$, we have*

$$\hat{R}_m(\mathcal{F}_S) \leqslant \inf_{0\leqslant\epsilon\leqslant c/2}\left\{4\epsilon + \frac{12}{\sqrt{m}}\int_{\epsilon}^{c/2}\sqrt{\ln C(\mathcal{F}, \nu, \|\cdot\|_{2,\mathcal{S}})}\,d\nu\right\}.$$

Proof: Fix

$$S = \{z_1, \ldots, z_m\}.$$

For each $j \in \mathbb{N}^+$, let

$$\epsilon_j = \frac{c}{2^j},$$

and let $\mathcal{F}_j$ be a minimal ϵ_j–cover of $\mathcal{F}$ with respect to the norm

$$\|f\|_{2,S} = \left(\frac{1}{m} \sum_{i=1}^{m} f^2(z_i) \right)^{1/2}.$$

Denote the covering number by

$$C_j = C(\mathcal{F}, \epsilon_j, \|\cdot\|_{2,S}).$$

For any $f \in \mathcal{F}$ and each $j \in \mathbb{N}^+$, choose

$$f_j \in \mathcal{F}_j \quad \text{such that} \quad \|f - f_j\|_{2,S} \leqslant \epsilon_j.$$

Then the sequence $\{f_j\}_{j \geqslant 1}$ converges (in the $\|\cdot\|_{2,S}$ metric) to f. This sequence allows us to write the telescoping (or chaining) decomposition

$$f = f_N + \sum_{j=1}^{N} (f_j - f_{j-1}), \qquad \text{with } f_0 = 0,$$

where $N \in \mathbb{N}$ is a parameter to be chosen later.

By the definition of the empirical Rademacher complexity we have

$$\hat{R}_m(\mathcal{F}_S) = \frac{1}{m} \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^{m} \sigma_i f(z_i).$$

Using the above telescoping sum we obtain

$$\hat{R}_m(\mathcal{F}_S) = \frac{1}{m} \mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \sum_{i=1}^{m} \sigma_i f_N(z_i) + \sum_{j=1}^{N} \sum_{i=1}^{m} \sigma_i (f_j(z_i) - f_{j-1}(z_i)) \right\}.$$

By the subadditivity of the supremum, we can split this into

$$\hat{R}_m(\mathcal{F}_S) \leqslant \frac{1}{m} \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^{m} \sigma_i f_N(z_i) + \sum_{j=1}^{N} \frac{1}{m} \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^{m} \sigma_i (f_j(z_i) - f_{j-1}(z_i)).$$

Bounding the first term. By the construction of the cover we have

$$\|f - f_N\|_{2,S} \leqslant \epsilon_N.$$

Hence, by a standard contraction argument,

$$\frac{1}{m} \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^{m} \sigma_i f_N(z_i)$$

can be made arbitrarily small by choosing N large enough (i.e. by taking ϵ_N sufficiently small). In our final bound this term will be absorbed by an additive 4ϵ term.

Bounding the chaining increments. For a fixed $j \in \{1, \ldots, N\}$, consider

$$T_j := \frac{1}{m}\, \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^{m} \sigma_i (f_j(z_i) - f_{j-1}(z_i)).$$

Since $f_j \in \mathcal{F}_j$ and $f_{j-1} \in \mathcal{F}_{j-1}$, there are at most $C_j C_{j-1}$ possible pairs (f_j, f_{j-1}). By Massart's Lemma (see, e.g., Theorem 4.3 in related texts), we have

$$T_j \leqslant \sqrt{\frac{2 \ln(C_j C_{j-1})}{m}} \cdot \sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^{m} (f_j(z_i) - f_{j-1}(z_i))^2 \right)^{1/2}.$$

Now, using the triangle inequality in $\|\cdot\|_{2,S}$,

$$\|f_j - f_{j-1}\|_{2,S} \leqslant \|f_j - f\|_{2,S} + \|f - f_{j-1}\|_{2,S} \leqslant \epsilon_j + \epsilon_{j-1}.$$

Since $\epsilon_{j-1} = \frac{c}{2^{j-1}} = 2\,\epsilon_j$, we have

$$\|f_j - f_{j-1}\|_{2,S} \leqslant 3\,\epsilon_j.$$

Thus,

$$T_j \leqslant 3\,\epsilon_j \sqrt{\frac{2 \ln(C_j C_{j-1})}{m}}.$$

For $j \geqslant 2$, the covering numbers are nonincreasing in ϵ, so $C_j \leqslant C_{j-1}$ and hence

$$\ln(C_j C_{j-1}) \leqslant 2 \ln C_j.$$

It follows that

$$T_j \leqslant \frac{6\,\epsilon_j}{\sqrt{m}} \sqrt{\ln C_j}.$$

Summing over $j = 1$ to N, we get

$$\sum_{j=1}^{N} T_j \leqslant \frac{6}{\sqrt{m}} \sum_{j=1}^{N} \epsilon_j \sqrt{\ln C_j}.$$

Converting the sum to an integral. Since $\epsilon_j = \frac{c}{2^j}$, the sum

$$\sum_{j=1}^{N} \epsilon_j \sqrt{\ln C_j}$$

can be viewed as a Riemann sum approximating the integral

$$\int_{\epsilon}^{c/2} \sqrt{\ln C(\mathcal{F}, \nu, \|\cdot\|_{2,S})}\, d\nu,$$

where $\epsilon > 0$ is chosen so that $\epsilon_{N+1} \leqslant \epsilon < \epsilon_N$. In particular, there is an absolute constant such that

$$\sum_{j=1}^{N} \epsilon_j \sqrt{\ln C_j} \leqslant 2 \int_{\epsilon}^{c/2} \sqrt{\ln C(\mathcal{F}, \nu, \|\cdot\|_{2,S})}\, d\nu.$$

Thus, the chaining increments are bounded by

$$\sum_{j=1}^{N} T_j \leqslant \frac{12}{\sqrt{m}} \int_{\epsilon}^{c/2} \sqrt{\ln C(\mathcal{F}, \nu, \|\cdot\|_{2,S})}\, d\nu.$$

Conclusion. Combining the bound on the first term with the bound on the chaining increments, we deduce that for any

$$0 \leqslant \epsilon \leqslant \frac{c}{2},$$

one has

$$\hat{R}_m(\mathcal{F}_S) \leqslant 4\epsilon + \frac{12}{\sqrt{m}} \int_{\epsilon}^{c/2} \sqrt{\ln C(\mathcal{F}, \nu, \|\cdot\|_{2,S})}\, d\nu.$$

Taking the infimum over $\epsilon \in [0, c/2]$ completes the proof. ■

References

[1] R. J. Liao, *Notes on Rademacher Complexity*, `https://www.cs.toronto.edu/~rjliao/notes/Notes_on_Rademacher_Complexity.pdf`. Accessed: 2025-03-24.

[2] Nina MF, *Lecture Notes on Machine Learning*, `https://www.cs.cmu.edu/~ninamf/ML11/lect1117.pdf`. Accessed: 2025-03-24.