Rademacher Complexity

Beier Zhu

Definition 1 (**Empirical Rademacher Complexity**). *Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{Z}$ to $[a,b]$ and $S=(z_1,...,z_m)$ a fixed sample of size m with elements in $\mathcal{Z}$. Then, the empirical Rademacher complexity of $\mathcal{G}$ w.r.t the sample S is defined as:*

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\frac{\boldsymbol{\sigma}^\top \mathbf{g}_S}{m}\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i)\right], \tag{1}$$

where $\boldsymbol{\sigma}=(\sigma_1,...,\sigma_m)^\top$, with σ_is independent uniform r.v.s taking values $\{-1,+1\}$. The r.v.s σ_i are called Rademacher variables. Let $\mathbf{g}_S$ denote the vector of values taken by function g over the sample S: $\mathbf{g}_S=(g(z_1),...,g(z_m))^\top$.

Remark.

- The inner product $\boldsymbol{\sigma}^\top\mathbf{g}_S$ measures the **correlation** of $\mathbf{g}_S$ with the vector of random noise $\boldsymbol{\sigma}$.
- The supremum $\sup_{g\in\mathcal{G}}\frac{\boldsymbol{\sigma}^\top\mathbf{g}_S}{m}$ measures **how well the function class** $\mathcal{G}$ correlates with $\boldsymbol{\sigma}$ over the sample S.
- The expectation on supremum $\mathbb{E}[\sup_{g\in\mathcal{G}}\frac{\boldsymbol{\sigma}^\top\mathbf{g}_S}{m}]$ measures **on average** how well the function class $\mathcal{G}$ correlates with noise on the sample S.

Definition 2 (**Rademacher Complexity**). *Let $\mathcal{D}$ denote the distribution from which samples are drawn. For any $m\geq 1$, the Rademacher complexity of $\mathcal{G}$ is the expectation of the empirical Rademacher complexity over all samples of size m drawn from $\mathcal{D}$:*

$$\mathfrak{R}_m(\mathcal{G}) = \mathbb{E}_{S\sim\mathcal{D}^m}\left[\widehat{\mathfrak{R}}_S(\mathcal{G})\right]. \tag{2}$$

Proposition 1. *For any $\delta>0$, with probability at least $1-\delta$, the following holds:*

$$\mathfrak{R}_m(\mathcal{G}) \leq \widehat{\mathfrak{R}}_S(\mathcal{G}) + (b-a)\sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{3}$$

Proof. Let S and S' be two samples differing by exactly one point, say z_m in S and z'_m in S'. Then, we have

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{G}) &= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\sigma_i g(z_i)\right] && (4)\\
&= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{m-1}\sigma_i g(z_i) + \sigma_m g(z'_m) + \sigma_m g(z_m) - \sigma_m g(z'_m)\right] && (5)\\
&\leq \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{m-1}\sigma_i g(z_i) + \sigma_m g(z'_m)\right] + \frac{1}{m}\mathbb{E}_{\sigma_m}\left[\sup_{g\in\mathcal{G}}\sigma_m\left(g(z_m)-g(z'_m)\right)\right] && (6)\\
&\leq \widehat{\mathfrak{R}}_{S'}(\mathcal{G}) + \frac{b-a}{m} && (7)
\end{aligned}$$

Similarly, we obtain $\widehat{\mathfrak{R}}_{S'}(\mathcal{G}) \leq \widehat{\mathfrak{R}}_S(\mathcal{G}) + \frac{b-a}{m}$, thus $|\widehat{\mathfrak{R}}_{S'}(\mathcal{G}) - \widehat{\mathfrak{R}}_S(\mathcal{G})| \leq \frac{b-a}{m}$. Then, we use McDiarmid's inequality to have Eq.(3). □

Theorem 1. *Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{Z}$ to $\mathbb{R}$. Let $\widehat{\mathbb{E}}_S[g(z)]$ denote the empirical average of g over S : $\widehat{\mathbb{E}}_S[g(z)] = \frac{1}{m}\sum_{i=1}^m g(z_i)$.*

$$\mathop{\mathbb{E}}_{S}\left[\sup_{g\in\mathcal{G}}\left[\widehat{\mathbb{E}}_S[g(z)] - \mathbb{E}[g(z)]\right]\right] \leq 2\mathfrak{R}_m(\mathcal{G}) \tag{8}$$

Proof. Fix $S = [z_1, ..., z_m]$, the term in the expectation on the LHS of Eq. (8) is

$$\sup_{g\in\mathcal{G}}\left[\widehat{\mathbb{E}}_S[g(z)] - \mathbb{E}[g(z)]\right] = \sup_{g\in\mathcal{G}}\left[\frac{1}{m}\sum_{i=1}^m g(z_i) - \mathbb{E}[g(z)]\right] \tag{9}$$

$$= \sup_{g\in\mathcal{G}}\left[\frac{1}{m}\sum_{i=1}^m g(z_i) - \mathbb{E}_{S'}[\frac{1}{m}\sum_{i=1}^m g(z_i')]\right] \qquad (\mathbb{E}[g] = \mathbb{E}_{S'}[\widehat{\mathbb{E}}_{S'}[g]])$$

$$= \frac{1}{m}\sup_{g\in\mathcal{G}}\left[\mathbb{E}_{S'}\left[\sum_{i=1}^m g(z_i) - g(z_i')\right]\right] \tag{10}$$

$$\leq \frac{1}{m}\mathbb{E}_{S'}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^m g(z_i) - g(z_i')\right] \qquad \text{(sub-add. of the sup.)}$$

Sub-additivity of the supremum functions: $\sup(U + V) \leq \sup(U) + \sup(V)$. Similarly, $\sup(\mathbb{E}_X[f(X)]) \leq \mathbb{E}_X[\sup(f(X))]$. Now, take the expectation over S for both side :

$$\mathop{\mathbb{E}}_{S}\left[\sup_{g\in\mathcal{G}}\left[\widehat{\mathbb{E}}_S[g(z)] - \mathbb{E}[g(z)]\right]\right] \leq \frac{1}{m}\mathop{\mathbb{E}}_{S,S'}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^m g(z_i) - g(z_i')\right] \tag{11}$$

$$= \frac{1}{m}\mathop{\mathbb{E}}_{S,S',\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^m \sigma_i\left(g(z_i) - g(z_i')\right)\right] \tag{12}$$

$$\leq \frac{1}{m}\mathop{\mathbb{E}}_{S,\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^m \sigma_i g(z_i)\right] + \frac{1}{m}\mathop{\mathbb{E}}_{S',\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^m -\sigma_i g(z_i')\right] \qquad (\sigma_i \stackrel{d}{=} -\sigma_i)$$

$$= 2\mathfrak{R}_m(\mathcal{G}) \tag{13}$$

Eq. (12) holds because $\sigma_i(g(z_i) - g(z_i')) \stackrel{d}{=} g(z_i) - g(z_i')$, since $g(z_i) - g(z_i')$ and $g(z_i') - g(z_i)$ have symmetric distribution. □

Theorem 2. *Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{Z}$ to $[0, 1]$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of an i.i.d. sample $S = [z_1, ...z_m]$ of size m, each of the following holds for all $g \in \mathcal{G}$:*

$$\mathbb{E}[g(z)] \leq \widehat{\mathbb{E}}_S[g(z)] + 2\mathfrak{R}_m(\mathcal{G}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{14}$$

$$\mathbb{E}[g(z)] \leq \widehat{\mathbb{E}}_S[g(z)] + 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}. \tag{15}$$

Compared to Theorem 2, we can rewrite the above as

$$\sup_{g\in\mathcal{G}}\left[\mathbb{E}[g(z)] - \widehat{\mathbb{E}}_S[g(z)]\right] \leq 2\mathfrak{R}_m(\mathcal{G}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{16}$$

Proof. The proof uses McDiarmid's inequality to function Φ defined for any sample S by

$$\Phi(S) = \sup_{g\in S}\left(\mathbb{E}[g] - \widehat{\mathbb{E}}_S[g]\right). \tag{17}$$

Let S and S' be two samples differing by exactly one point, say z_m in S and z'_m in S'. Then, since the difference of suprema does not exceed the supremum of the difference, we have

$$\Phi(S') - \Phi(S) \leq \sup_{g \in \mathcal{G}} \left(\widehat{\mathbb{E}}_S[g] - \widehat{\mathbb{E}}_{S'}[g] \right) = \sup_{g \in \mathcal{G}} \frac{g(z_m) - g(z'_m)}{m} \leq \frac{1}{m}. \tag{18}$$

Similarly, we obtain $\Phi(S) - \Phi(S') \leq \frac{1}{m}$, thus $|\Phi(S) - \Phi(S')| \leq \frac{1}{m}$. Then, by McDiarmid's inequality, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds:

$$\Phi(S) \leq \mathop{\mathbb{E}}_S[\Phi(S)] + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}. \tag{19}$$

We next bound the expectation of the right-hand side as follows:

$$\begin{aligned}
\mathop{\mathbb{E}}_S[\Phi(S)] &= \mathop{\mathbb{E}}_S \left[\sup_{g \in \mathcal{G}} (\mathbb{E}(g) - \widehat{\mathbb{E}}_S(g)) \right] && (20) \\
&= \mathop{\mathbb{E}}_S \left[\sup_{g \in \mathcal{G}} \mathop{\mathbb{E}}_{S'} [\widehat{\mathbb{E}}_{S'}(g) - \widehat{\mathbb{E}}_S(g)] \right] && (\mathbb{E}[g] = \mathbb{E}_{S'}[\widehat{\mathbb{E}}_{S'}[g]]) \\
&\leq \mathop{\mathbb{E}}_{S,S'} \left[\sup_{g \in \mathcal{G}} (\widehat{\mathbb{E}}_{S'}(g) - \widehat{\mathbb{E}}_S(g)) \right] && \text{(sub-add. of the sup.)} \\
&= \mathop{\mathbb{E}}_{S,S'} \left[\frac{1}{m} \sum_{i=1}^{m} \sup_{g \in \mathcal{G}} g(z'_i) - g(z_i) \right] && (21) \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\sigma},S,S'} \left[\frac{1}{m} \sum_{i=1}^{m} \sup_{g \in \mathcal{G}} \sigma_i (g(z'_i) - g(z_i)) \right] && (22) \\
&\leq \mathop{\mathbb{E}}_{\boldsymbol{\sigma},S'} \left[\frac{1}{m} \sum_{i=1}^{m} \sup_{g \in \mathcal{G}} \sigma_i g(z'_i) \right] + \mathop{\mathbb{E}}_{\boldsymbol{\sigma},S} \left[\frac{1}{m} \sum_{i=1}^{m} \sup_{g \in \mathcal{G}} -\sigma_i g(z_i) \right] && \text{(sub-add. of the sup.)} \\
&= 2\mathfrak{R}_m(\mathcal{G}) && (-\sigma \text{ is Rademacher r.v.})
\end{aligned}$$

In Eq.(22), we introduce Rademacher variables σ_i which do not change the expectation appearing in Eq.(21): when $\sigma_i = 1$, the associated summand remains unchanged; when $\sigma_i = -1$, the associated summand flips signs, which equivalent to swapping z_i and z'_i between S and S'. Since (z_i, z'_i) and (z'_i, z_i) have the same joint distribution, this swap does not affect the overall expectation. Substituting it back into Eq. (19) and using the definition of $\Phi(S)$, we obtain

$$\sup_{g \in S} \left(\mathbb{E}[g] - \widehat{\mathbb{E}}_S[g] \right) \leq 2\mathfrak{R}_m(\mathcal{G}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{23}$$

Obviously, $\forall g \in \mathcal{G}$, we have:

$$\mathbb{E}[g(z)] \leq \widehat{\mathbb{E}}_S[g(z)] + 2\mathfrak{R}_m(\mathcal{G}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{24}$$

Using Proposition 1, we have $\mathfrak{R}_m(\mathcal{G}) \leq \widehat{\mathfrak{R}}_S(\mathcal{G}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$. Then,

$$\mathbb{E}[g(z)] \leq \widehat{\mathbb{E}}_S[g(z)] + 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{25}$$

□

Lemma 1. *Let $\mathcal{H}$ be a family of functions taking values in $\{-1, +1\}$ and let $\mathcal{G}$ be the family of loss functions associated to $\mathcal{H}$ for the zero-one loss: $\mathcal{G} = \{(x, y) \mapsto \mathbb{1}[h(x) \neq y] : h \in \mathcal{H}\}$. For any sample $S = ((x_1, y_1), ..., (x_m, y_m))$ of elements in $\mathcal{X} \times \{-1, +1\}$, let $S_{\mathcal{X}}$ denote its projection over $\mathcal{X} : S_{\mathcal{X}} = (x_1, ..., x_m)$. Then, the following relation holds between the empirical Rademacher complexities of $\mathcal{G}$ and $\mathcal{H}$:*

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \frac{1}{2} \widehat{\mathfrak{R}}_{S_{\mathcal{X}}}(\mathcal{H}). \tag{26}$$

Proof.

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i \mathbb{1}[h(x_i)\neq y_i]\right] \tag{27}$$

$$= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i \frac{1-h(x_i)y_i}{2}\right] \tag{28}$$

$$= \frac{1}{2}\frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}(-y_i\sigma_i)h(x_i)\right] \tag{29}$$

$$= \frac{1}{2}\widehat{\mathfrak{R}}_{S_{\mathcal{X}}}(\mathcal{H}), \tag{30}$$

where we use the fact that $\mathbb{1}[h(x_i)\neq y_i] = \frac{1-h(x_i)y_i}{2}$ and the fact that for fixed $y_i = \{-1,+1\}$, $(-\sigma_i y_i)$ are also Rademacher r.v.s. □

Theorem 3. *Let $\mathcal{H}$ be a family of functions taking values in $\{-1,+1\}$ and let $\mathcal{D}$ be the distribution over the input space $\mathcal{X}$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over a sample S of size m drawn according to $\mathcal{D}$, each of the following holds for any $h\in\mathcal{H}$:*

$$\mathcal{R}(h) \leq \widehat{\mathcal{R}}_S(h) + \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{31}$$

$$\mathcal{R}(h) \leq \widehat{\mathcal{R}}_S(h) + \widehat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{32}$$

Proof. The result follows immediately by Theorem 2 and Lemma 1. □

Remark. The second bound is data-dependent: the empirical Rademacher complexity $\widehat{\mathfrak{R}}_S(\mathcal{H})$ is a function of a specific sample S.

Proposition 2 (Talagrand's Lemma). *Let $\Phi_1,...,\Phi_m$ be l-Lipschitz functions from $\mathbb{R}$ to $\mathbb{R}$ and $\sigma_1,...,\sigma_m$ be Rademacher variable. Then, for any hypothesis set $\mathcal{H}$ of real-valued functions, the following inequality holds:*

$$\frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i(\Phi_i\circ h)(x_i)\right] \leq \frac{l}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i h(x_i)\right] = l\widehat{\mathfrak{R}}_S(\mathcal{H}). \tag{33}$$

In particular, if $\Phi_i = \Phi$ for all $i\in[m]$, then the following holds:

$$\widehat{\mathfrak{R}}_S(\Phi\circ\mathcal{H}) \leq l\widehat{\mathfrak{R}}_S(\mathcal{H}). \tag{34}$$

Proof. First we fix a sample $S = (x_1,...x_m)$, then, by definition,

$$\frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i(\Phi_i\circ h)(x_i)\right] = \frac{1}{m}\mathbb{E}_{\sigma_{1:m-1}}\left[\mathbb{E}_{\sigma_m}\left[\sup_{h\in\mathcal{H}} u_{m-1}(h) + \sigma_m(\Phi_m\circ h)(x_m)\right]\right], \tag{35}$$

where $u_{m-1}(h) = \sum_{i=1}^{m-1} = \sigma_i(\Phi_i\circ h)(x_i)$. By the definition of the supremum ($s = \sup A$, then for any $\epsilon > 0$, $\exists a_\epsilon \in A$ s.t. $a_\epsilon > s-\epsilon$.), for any $\epsilon > 0$, there exist $h_1, h_2\in\mathcal{H}$ such that

$$u_{m-1}(h_1) + (\Phi_m\circ h_1)(x_m) \geq (1-\epsilon)\left[\sup_{h\in\mathcal{H}} u_{m-1}(h) + (\Phi_m\circ h)(x_m)\right] \tag{36}$$

$$u_{m-1}(h_2) - (\Phi_m\circ h_2)(x_m) \geq (1-\epsilon)\left[\sup_{h\in\mathcal{H}} u_{m-1}(h) - (\Phi_m\circ h)(x_m)\right] \tag{37}$$

Thus, for any $\epsilon > 0$, by definition of $\mathbb{E}_{\sigma_m}$,

$$(1-\epsilon)\mathop{\mathbb{E}}_{\sigma_m}\left[\sup_{h\in\mathcal{H}} u_{m-1}(h) + \sigma_m(\Phi_m \circ h)(x_m)\right] \tag{38}$$

$$=(1-\epsilon)\left[\frac{1}{2}\sup_{h\in\mathcal{H}}\left[u_{m-1}(h) + (\Phi_m \circ h)(x_m)\right] + \frac{1}{2}\sup_{h\in\mathcal{H}}\left[u_{m-1}(h) - (\Phi_m \circ h)(x_m)\right]\right] \tag{39}$$

$$\leq\frac{1}{2}\left[u_{m-1}(h_1) + (\Phi_m \circ h_1)(x_m)\right] + \frac{1}{2}\left[u_{m-1}(h_2) - (\Phi_m \circ h_2)(x_m)\right] \tag{40}$$

Let $s = \text{sign}(h_1(x_m) - h_2(x_m))$. Then, the previous inequality implies:

$$=\frac{1}{2}\left[u_{m-1}(h_1) + u_{m-1}(h_2) + \Phi_m(h_1(x_m)) - \Phi_m(h_2(x_m))\right] \qquad \text{(rearranging)}$$

$$\leq\frac{1}{2}\left[u_{m-1}(h_1) + u_{m-1}(h_2) + sl(h_1(x_m) - h_2(x_m))\right] \qquad \text{(l-Lipschitzness)}$$

$$=\frac{1}{2}\left[u_{m-1}(h_1) + slh_1(x_m)\right] + \frac{1}{2}\left[u_{m-1}(h_2) - slh_2(x_m)\right] \qquad \text{(rearranging)}$$

$$\leq\frac{1}{2}\sup_{h\in\mathcal{H}}\left[u_{m-1}(h) + slh(x_m)\right] + \frac{1}{2}\sup_{h\in\mathcal{H}}\left[u_{m-1}(h) - slh(x_m)\right] \qquad \text{(definition of sup)}$$

$$=\mathop{\mathbb{E}}_{\sigma_m}\left[\sup_{h\in\mathcal{H}} u_{m-1}(h) + \sigma_m lh(x_m)\right] \qquad \text{(definition of } \mathop{\mathbb{E}}_{\sigma_m}\text{)}$$

Since the inequality holds for all $\epsilon > 0$, we have:

$$\mathop{\mathbb{E}}_{\sigma_m}\left[\sup_{h\in\mathcal{H}} u_{m-1}(h) + \sigma_m(\Phi_m \circ h)(x_m)\right] \leq \mathop{\mathbb{E}}_{\sigma_m}\left[\sup_{h\in\mathcal{H}} u_{m-1}(h) + \sigma_m lh(x_m)\right] \tag{41}$$

Proceeding in the same way for all other $\sigma_i (i \neq m)$ proves the lemma. □

Proposition 3 (Extending Talagrand's Lemma to Vector Valued Functions). *Let $\mathcal{H}$ be a hypothesis set of functions mapping $\mathcal{X}$ to $\mathbb{R}^c$. Assume that for all $i = 1, ..., m$, $\Psi_i : \mathbb{R}^c \to \mathbb{R}$ is μ_i-Lipschitz for $\mathbb{R}^c$ equipped with the 2-norm. That is:*

$$|\Psi_i(\mathbf{x}') - \Psi_i(\mathbf{x})| \leq \|\mathbf{x}' - \mathbf{x}\|_2, \tag{42}$$

for all $(\mathbf{x}, \mathbf{x}') \in (\mathbb{R}^c, \mathbb{R}^c)$. Then, for any sample S of m points $x_1, ..., x_m \in \mathcal{X}$, the following inequality holds

$$\frac{1}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{h}\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i\Psi_i(\mathbf{h}(x_i))\right] \leq \frac{\sqrt{2}}{m}\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}}\left[\sup_{\mathbf{h}\in\mathcal{H}}\sum_{i=1}^{m}\sum_{j=1}^{c}\epsilon_{ij}\mu_i h_j(x_i)\right], \tag{43}$$

where $\boldsymbol{\epsilon} = (\epsilon_{ij})_{i,j}$ and ϵ_{ij} are independent Rademacher variables uniformly distributed over $\{1, -1\}$. In particular, if $\Psi_i = \Psi$ for all $i \in [m]$, then the following holds:

$$\widehat{\mathfrak{R}}_S(\Psi \circ \mathcal{H}) \leq \frac{\sqrt{2}}{m}\mu\widehat{\mathfrak{R}}_S(\mathcal{H}), \tag{44}$$

Proof. First we fix a sample $S = (x_1, ..., x_m)$, then, by definition,

$$\frac{1}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{h}\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i\Psi_i(\mathbf{h}(x_i))\right] = \frac{1}{m}\mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_{m-1}}\left[\mathop{\mathbb{E}}_{\sigma_m}\left[\sup_{\mathbf{h}\in\mathcal{H}} U_{m-1}(\mathbf{h}) + \sigma_m\Psi_m(\mathbf{h}(x_m))\right]\right], \tag{45}$$

where $U_{m-1}(\mathbf{h}) = \sum_{i=1}^{m-1}\sigma_i\Psi_i(\mathbf{h}(x_i))$. Assume that the suprema can be attained and let $\mathbf{h}_1, \mathbf{h}_2 \in \mathcal{H}$ be the hypotheses satisfying

$$U_{m-1}(\mathbf{h}_1) + \Psi_m(\mathbf{h}_1(x_m)) = \sup_{\mathbf{h}\in\mathcal{H}} U_{m-1}(\mathbf{h}) + \Psi_m(\mathbf{h}(x_m)) \tag{46}$$

$$U_{m-1}(\mathbf{h}_2) - \Psi_m(\mathbf{h}_2(x_m)) = \sup_{\mathbf{h}\in\mathcal{H}} U_{m-1}(\mathbf{h}) - \Psi_m(\mathbf{h}(x_m)) \tag{47}$$

When the suprema are not reached, a similar argument to what follows can be given by considering instead hypotheses that are ϵ-close to the suprema for any $\epsilon > 0$. By definition of expectation, since σ_m is uniformly distributed over $\{1, -1\}$, we can write

$$\mathbb{E}_{\sigma_m}\left[\sup_{\mathbf{h}\in\mathcal{H}} U_{m-1}(\mathbf{h}) + \sigma_m \Psi_m(\mathbf{h}(x_m))\right] \tag{48}$$

$$= \frac{1}{2}\sup_{\mathbf{h}\in\mathcal{H}} U_{m-1}(\mathbf{h}) + \Psi_m(\mathbf{h}(x_m)) + \frac{1}{2}\sup_{\mathbf{h}\in\mathcal{H}} U_{m-1}(\mathbf{h}) - \Psi_m(\mathbf{h}(x_m)) \tag{49}$$

$$= \frac{1}{2}\left[U_{m-1}(\mathbf{h}_1) + \Psi_m(\mathbf{h}_1(x_m))\right] + \frac{1}{2}\left[U_{m-1}(\mathbf{h}_2) - \Psi_m(\mathbf{h}_2(x_m))\right] \tag{50}$$

$$= \frac{1}{2}\left[U_{m-1}(\mathbf{h}_1) + U_{m-1}(\mathbf{h}_2) + \Psi_m(\mathbf{h}_1(x_m)) - \Psi_m(\mathbf{h}_2(x_m))\right] \tag{51}$$

$$\leq \frac{1}{2}\left[U_{m-1}(\mathbf{h}_1) + U_{m-1}(\mathbf{h}_2) + \mu_m \|\mathbf{h}_1(x_m) - \mathbf{h}_2(x_m)\|_2\right] \tag{52}$$

$$\leq \frac{1}{2}\left[U_{m-1}(\mathbf{h}_1) + U_{m-1}(\mathbf{h}_2) + \mu_m\sqrt{2}\mathbb{E}_{\epsilon_{m1},\ldots,\epsilon_{mc}}\left[\left|\sum_{j=1}^{c}\epsilon_{mj}(h_{1j}(x_m) - h_{2j}(x_m))\right|\right]\right], \tag{53}$$

where we use the μ_m-Lipschitzness of Ψ_m and the Khintchine-Kahane inequality. Let $\boldsymbol{\epsilon}_m = (\epsilon_{m1}, ..., \epsilon_{mc})$ and $s(\boldsymbol{\epsilon}_m) \in \{1, -1\}$ denote the sign of $\sum_{j=1}^{c}\epsilon_{mj}(h_{1j}(x_m) - h_{2j}(x_m))$. Then, the following holds:

$$\mathbb{E}_{\sigma_m}\left[\sup_{\mathbf{h}\in\mathcal{H}} U_{m-1}(\mathbf{h}) + \sigma_m \Psi_m(\mathbf{h}(x_m))\right] \tag{54}$$

$$\leq \frac{1}{2}\mathbb{E}_{\boldsymbol{\epsilon}_m}\left[U_{m-1}(\mathbf{h}_1) + U_{m-1}(\mathbf{h}_2) + \mu_m\sqrt{2}\left[\left|\sum_{j=1}^{c}\epsilon_{mj}(h_{1j}(x_m) - h_{2j}(x_m))\right|\right]\right] \tag{55}$$

$$= \frac{1}{2}\mathbb{E}_{\boldsymbol{\epsilon}_m}\left[U_{m-1}(\mathbf{h}_1) + \mu_m\sqrt{2}s(\boldsymbol{\epsilon}_m)\sum_{j=1}^{c}\epsilon_{mj}h_{1j}(x_m) + U_{m-1}(\mathbf{h}_2) - \mu_m\sqrt{2}s(\boldsymbol{\epsilon}_m)\sum_{j=1}^{c}\epsilon_{mj}h_{2j}(x_m)\right] \tag{56}$$

$$\leq \frac{1}{2}\mathbb{E}_{\boldsymbol{\epsilon}_m}\left[\sup_{\mathbf{h}\in\mathcal{H}}\left(U_{m-1}(\mathbf{h}) + \mu_m\sqrt{2}s(\boldsymbol{\epsilon}_m)\sum_{j=1}^{c}\epsilon_{mj}h_j(x_m)\right) + \sup_{\mathbf{h}\in\mathcal{H}}\left(U_{m-1}(\mathbf{h}) - \mu_m\sqrt{2}s(\boldsymbol{\epsilon}_m)\sum_{j=1}^{c}\epsilon_{mj}h_j(x_m)\right)\right] \tag{57}$$

$$= \mathbb{E}_{\boldsymbol{\epsilon}_m}\left[\mathbb{E}_{\sigma_m}\left[\sup_{\mathbf{h}\in\mathcal{H}} U_{m-1}(\mathbf{h}) + \mu_m\sqrt{2}\sigma_m\sum_{j=1}^{c}\epsilon_{mj}h_j(x_m)\right]\right] \tag{58}$$

$$= \mathbb{E}_{\boldsymbol{\epsilon}_m}\left[\sup_{\mathbf{h}\in\mathcal{H}} U_{m-1}(\mathbf{h}) + \mu_m\sqrt{2}\sum_{j=1}^{c}\epsilon_{mj}h_j(x_m)\right] \tag{59}$$

We have the last equality as the product of two independent Rademacher variables ($\sigma_m\epsilon_{mj}$) is still Rademacher variable. Note that $\mathbb{E}[\epsilon_i\epsilon_j a] = 0$ for fixed a, but $\mathbb{E}[\sup_{a\in\mathcal{A}}\epsilon_i\epsilon_j a] \neq 0$. For example, if $\mathcal{A} = \{1, -1\}$, $\mathbb{E}[\sup_{a\in\mathcal{A}}\epsilon_i\epsilon_j a] = 1$. Proceeding in the same way for all other σ_is ($i < m$) completes the proof.

□

Rademacher Identities

Fix $m \geq 1$, for any $\alpha \in \mathbb{R}$ and any two hypothesis sets $\mathcal{H}$ and $\mathcal{H}'$ of functions mapping from $\mathcal{X}$ to $\mathbb{R}$, we have:

(a) $\mathfrak{R}_m(\alpha\mathcal{H}) = |\alpha|\mathfrak{R}_m(\mathcal{H})$,
where $\alpha\mathcal{H} = \{\alpha h(x) | h \in \mathcal{H}\}$.

(b) $\mathfrak{R}_m(\alpha + \mathcal{H}) = \mathfrak{R}_m(\mathcal{H})$,
where $\alpha + \mathcal{H} = \{\alpha + h(x) | h \in \mathcal{H}\}$.

(c) $\mathfrak{R}_m(\mathcal{H} + \mathcal{H}') = \mathfrak{R}_m(\mathcal{H}) + \mathfrak{R}_m(\mathcal{H}')$,
where $\mathcal{H} + \mathcal{H}' = \{h(x) + h'(x) | h \in \mathcal{H}, h \in \mathcal{H}'\}$.

(d) $\mathfrak{R}_m(\{\max(h, h') | h \in \mathcal{H}, h \in \mathcal{H}'\}) \leq \mathfrak{R}_m(\mathcal{H}) + \mathfrak{R}_m(\mathcal{H}')$.
$\max(h, h') : x \mapsto \max(h(x), h'(x))$.

Fix $\mathbf{x} \in \mathbb{R}^m$ and let $\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_m)^\top$ with σ_is be Rademacher variables, Then:

(e) $\|\mathbf{x}\|_2 = \left[\mathbb{E}_{\boldsymbol{\sigma}}(\boldsymbol{\sigma}^\top \mathbf{x})^2\right]^{\frac{1}{2}}$.

(f) Khintchine inequality.

$$A_p \|\mathbf{x}\|_2 \leq \left[\mathbb{E}_{\boldsymbol{\sigma}} |\boldsymbol{\sigma}^\top \mathbf{x}|^p\right]^{1/p} \leq B_p \|\mathbf{x}\|_2$$

where

$$A_p = \begin{cases} 2^{1/2-1/p} & \text{if } 0 < p \leq p_0, \\ 2^{1/2} \left(\Gamma\left(\frac{p+1}{2}\right) / \sqrt{\pi}\right)^{1/p} & \text{if } p_0 < p < 2, \\ 1 & \text{if } 2 \leq p < \infty, \end{cases} \tag{60}$$

$$\text{and} \quad B_p = \begin{cases} 1 & \text{if } 0 < p \leq 2, \\ 2^{1/2} \left(\Gamma\left(\frac{p+1}{2}\right) / \sqrt{\pi}\right)^{1/p} & \text{if } 2 < p < \infty. \end{cases} \tag{61}$$

$p_0 \approx 1.847$ and Γ is the Gamma function.

Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two families of functions mapping $\mathcal{X}$ to $\{0, 1\}$ and let $\mathcal{H} = \{h_1 h_2 : h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$, Then:

(g) $\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq \widehat{\mathfrak{R}}_S(\mathcal{H}_1) + \widehat{\mathfrak{R}}_S(\mathcal{H}_2)$

Proof. We proof the equalities for empirical Rademacher complexity over a sample S, then taking the expectation yields the claimed results.

(a) For fixed sample S, we have

$$\widehat{\mathfrak{R}}_S(\alpha\mathcal{H}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \sigma_i \alpha h(x_i)\right] = |\alpha| \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \sigma_i \text{sign}(\alpha) h(x_i)\right] = |\alpha| \widehat{\mathfrak{R}}_S(\mathcal{H}). \tag{62}$$

Note that $\sigma_i \text{sign}(\alpha)$s are still Rademacher variables and $\sup cA = c \sup A$ if $c \geq 0$.

(b) For fixed sample S, we have

$$\widehat{\mathfrak{R}}_S(\alpha + \mathcal{H}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \sigma_i [\alpha + h(x_i)]\right] \tag{63}$$

$$= \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{m}\alpha\sum_{i=1}^m \sigma_i\right] + \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \sigma_i h(x_i)\right] \tag{64}$$

$$= \widehat{\mathfrak{R}}_S(\mathcal{H}) \qquad (\mathbb{E}_{\boldsymbol{\sigma}}[\textstyle\sum_i \sigma_i] = 0)$$

(c) For fixed sample S, we have

$$\widehat{\mathfrak{R}}_S(\mathcal{H} + \mathcal{H}') = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}, h'\in\mathcal{H}'} \frac{1}{m}\sum_{i=1}^m \sigma_i \left(h(x_i) + h'(x_i)\right)\right] \tag{65}$$

$$= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \sigma_i \left(h(x_i)\right) + \sup_{h\in\mathcal{H}'} \frac{1}{m}\sum_{i=1}^m \sigma_i \left(h'(x_i)\right)\right] \tag{66}$$

$$= \widehat{\mathfrak{R}}_S(\mathcal{H}) + \widehat{\mathfrak{R}}_S(\mathcal{H}') \tag{67}$$

(d) Fix sample S and use the identity $\max(a,b) = \frac{1}{2}[a+b+|a-b|]$, we have:

$$\widehat{\mathfrak{R}}_S(\max(\mathcal{H},\mathcal{H}')) \tag{68}$$

$$= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H},h'\in\mathcal{H}'} \frac{1}{m}\sum_{i=1}^{m}\sigma_i \max(h(x_i), h'(x_i))\right] \tag{69}$$

$$= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H},h'\in\mathcal{H}'} \frac{1}{m}\sum_{i=1}^{m}\sigma_i \frac{1}{2}\left(h(x_i)+h'(x_i)+|h(x_i)-h'(x_i)|\right)\right] \tag{70}$$

$$= \frac{1}{2}\left[\widehat{\mathfrak{R}}_S(\mathcal{H})+\widehat{\mathfrak{R}}_S(\mathcal{H}')\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h,h'} \frac{1}{m}\sum_{i=1}^{m}\sigma_i|h(x_i)-h'(x_i)|\right] \quad \text{(using Eq.(67))}$$

$$\leq \frac{1}{2}\left[\widehat{\mathfrak{R}}_S(\mathcal{H})+\widehat{\mathfrak{R}}_S(\mathcal{H}')\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h,h'} \frac{1}{m}\sum_{i=1}^{m}\sigma_i h(x_i)-\sigma_i h'(x_i)\right] \quad (\phi(t)=|t| \text{ is 1-Lipschitz})$$

$$= \frac{1}{2}\left[\widehat{\mathfrak{R}}_S(\mathcal{H})+\widehat{\mathfrak{R}}_S(\mathcal{H}')\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h} \frac{1}{m}\sum_{i=1}^{m}\sigma_i h(x_i)\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h'} \frac{1}{m}\sum_{i=1}^{m}(-\sigma_i) h'(x_i)\right] \tag{71}$$

$$= \widehat{\mathfrak{R}}_S(\mathcal{H})+\widehat{\mathfrak{R}}_S(\mathcal{H}') \quad (-\sigma_i\text{s are Rademacher r.v.s})$$

(e) Recall that $\mathbb{E}[\sigma_i\sigma_j]=0$ if $i\neq j$ and $\mathbb{E}[\sigma_i^2]=1$, we have

$$\left[\mathbb{E}_{\boldsymbol{\sigma}}(\boldsymbol{\sigma}^\top \mathbf{x})^2\right]^{\frac{1}{2}} = \left[\mathbb{E}_{\boldsymbol{\sigma}}(\sum_i \sigma_i x_i)^2\right]^{\frac{1}{2}} = \left[\sum_i \mathbb{E}[\sigma_i^2]x_i^2 + 2\sum_{i\neq j}\mathbb{E}[\sigma_i\sigma_j]x_i x_j\right]^{\frac{1}{2}} = \|\mathbf{x}\|_2. \tag{72}$$

(g) Note that for $a,b\in\{0,1\}^2$, we have $ab=\max(0,a+b-1)$. Then,

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) = \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h_1\in\mathcal{H}_1,h_2\in\mathcal{H}_2}\sum_{i=1}^{m}\sigma_i h_1(x_i)h_2(x_i)\right] \tag{73}$$

$$= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h_1\in\mathcal{H}_1,h_2\in\mathcal{H}_2}\sum_{i=1}^{m}\sigma_i \max(0, h_1(x_i)+h_2(x_i)-1)\right] \tag{74}$$

Let $g(t)=\max(0,t-1)$ which is 1-Lipschitz. Using Talagrand's Lemma, we have:

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h_1\in\mathcal{H}_1,h_2\in\mathcal{H}_2}\sum_{i=1}^{m}\sigma_i (h_1(x_i)+h_2(x_i))\right] \tag{75}$$

$$= \widehat{\mathfrak{R}}_S(\mathcal{H}_1)+\widehat{\mathfrak{R}}_S(\mathcal{H}_2) \tag{76}$$

□