# Machine Learning Theory Lecture 3: Rademacher Complexity

Richard Xu

April 30, 2022

## 1 Definition

let each of $S = \{Z_i\}$ be distributed from a data distribution $\mathcal{D}$

$$\text{Rad}_n(\mathcal{H}) = \mathbb{E}_S\Big[\mathbb{E}_{\bar{\sigma}}\Big[\sup_{h\in\mathcal{H}} \frac{\sum_{i=1}^n \sigma_i h(Z_i)}{n}\Big]\Big] \tag{1}$$

1. In words, We sample $n$ data $\{Z_i\}_{i=1}^n$ at random from $\mathcal{D}$; We also sample $n$ random binary labels from Radmarcher distribution. What is the "average of the best correlations" can hypothesis set $\mathcal{H}$ achieve? Obviously, the higher the correlations that $h \in \mathcal{H}$ can achieve between the set $\{Z_i\}_{i=1}^n$ and the set $\{\sigma_i\}_{i=1}^n$, a better performance (or complexity) for $\mathcal{H}$.

2. Obviously, the most difficult for computing $\text{Rad}_n(\mathcal{H})$ is to $\max$ over a possibly infinite hypothesis set $\mathcal{H}$ (for example all the lines in linear classifications). Lucikly, we can take advantage of for example:

   (a) there are infinite $h$, but $h(Z_i)$ has finite outcome.

   (b) or the algebraic property, for example: when $h(\boldsymbol{w}^\top \mathbf{x}) = \sup_w(w^\top \mathbf{x})$

### 1.1 alternative definition

however, some text are using definition:

$$\text{Rad}_n(\mathcal{H}) = \mathbb{E}_S\Big[\mathbb{E}_{\bar{\sigma}}\Big[\sup_{h\in\mathcal{H}} \Big|\frac{\sum_{i=1}^n \sigma_i h(Z_i)}{n}\Big|\Big]\Big] \tag{2}$$

**QUESTION** is the above definition also valid?

## 1.2 Empirical Rademacher Complexity

$$\widehat{\text{Rad}}_S(\mathcal{H}) = \mathbb{E}_{\bar{\sigma}}\Big[\sup_{h\in\mathcal{H}} \frac{\sum_{i=1}^{n} \sigma_i h(Z_i)}{n}\Big] \tag{3}$$

which is precisely the stuff inside $\text{Rad}_n(\mathcal{H})$, i.e.,

$$\text{Rad}_n(\mathcal{H}) = \mathbb{E}_S\Big[\widehat{\text{Rad}}_S(\mathcal{H})\Big] \tag{4}$$

# 2 Bounds

## 2.1 Expected hypothesis bound

Before we state the expression for the generalization error bound in terms of expected and empirical loss, let's begin with a generic bound in terms of Rademarcher complexity:

**Theorem 1** *Let $Z, Z_1, \ldots, Z_n$ be i.i.d random variables sampled from $\mathcal{D}$, and consider every hypothesis $f \in \mathcal{F}$ is bounded by $[a, b]$. Then, $\forall\delta > 0$, with probability of at least $1-\delta$, and respect to sample $S$, we have:*

$$\forall f \in \mathcal{F}: \quad \mathbb{E}_Z[f(Z)] \leq \frac{1}{n}\sum_{i=1}^{n} f(Z_i) + 2Rad_n(\mathcal{F}) + (b-a)\sqrt{\frac{\log(1/\delta)}{2n}} \tag{5}$$

$$\forall f \in \mathcal{F}: \quad \mathbb{E}_Z[f(Z)] \leq \frac{1}{n}\sum_{i=1}^{n} f(Z_i) + 2\widehat{Rad}_S(\mathcal{F}) + 3(b-a)\sqrt{\frac{\log(2/\delta)}{2n}} \tag{6}$$

Please note the following:

1. some literature write it as:

$$\forall f \in \mathcal{F}: \quad \mathbb{E}_Z[f(Z)] \leq \frac{1}{n}\sum_{i=1}^{n} f(Z_i) + 2\text{Rad}_n(\mathcal{F}) + (b-a)\sqrt{\frac{\log(1/\delta)}{n}} \tag{7}$$

   which is also valid.

2. note how I deliberately write it using $f \in \mathcal{F}$ instead of $h \in \mathcal{H}$ because we have:

$$f_h(Z_i) = \ell(h(\mathbf{x}_i), y_i) \tag{8}$$

3. note that Theorem (1) can be applied more generically, and it does not need to apply specifically for Generalization error bound (in terms of expected and empirical loss) as stated in Theorem generalization bound rademarcher complexity. In order to reflect this, let's give a symbol for:

$$\begin{aligned} T_n &= \mathbb{E}_Z[f(Z)] \\ \widehat{T}_S &= \frac{1}{n}\sum_{i=1}^{n} f(Z_i) \end{aligned} \tag{9}$$

4. I will show the proof of Theorem (1) later using McDiarmid inequality. in Section (3)

## 2.2 Generalization error bound

Firstly, Generalization error bound is what we are interested in. The expected and empirical risks are expressed as:

$$\begin{aligned} R(h) &= \mathbb{E}_Z[f_h(Z)] \\ &= \mathbb{E}_Z[\ell(h(\mathbf{x}), y)] \\ \hat{R}_S(h) &= \frac{1}{n}\sum_{i=1}^{n} f_h(Z_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i) \end{aligned} \tag{10}$$

**Theorem 2** *let $\mathcal{H}$ be set of hypothesis taking values in $\{-1, +1\}$ and $\mathcal{F} = L(\mathcal{H}) = \{f_h(x, y) \rightarrow \mathbb{1}_{\{h(x)\neq y\}} : h \in \mathcal{H}\}$. For any $\delta > 0$, with probability at least $1 - \delta$ over a sample $S$ of size $n$ drawn from $\mathcal{D}$:*

$$\forall h \in \mathcal{H}: \quad R(h) \leq \hat{R}_S(h) + \widehat{Rad}_S(\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2n}} \tag{11}$$

note that we need to apply Rademacher calculus to $\mathcal{H}$ taking values in other ranges.

## 2.3 first attempt

Now, a straight application of Theorem 1 will give the following expression, let $a = -1$, $b = 1$, and let $f(Z) \equiv f_h(Z)$:

$$\begin{aligned} &\forall h \in \mathcal{H}: \quad \mathbb{E}_Z[f_h(Z)] \leq \frac{1}{n}\sum_{i=1}^{n} f_h(Z_i) + 2\widehat{\text{Rad}}_S(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}} \\ \implies &\forall h \in \mathcal{H}: \quad R(h) \leq \hat{R}_S(h) + 2\widehat{\text{Rad}}_S(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}} \\ &\qquad\qquad = \hat{R}_S(h) + 2\widehat{\text{Rad}}_S(L(\mathcal{H})) + 3\sqrt{\frac{\log(2/\delta)}{2n}} \end{aligned} \tag{12}$$

where $\mathcal{F} = L(\mathcal{H}) = \{f_h(x, y) \rightarrow \mathbb{1}_{\{h(x)\neq y\}} : h \in \mathcal{H}\}$

Note that we do not like to express it in terms of $\widehat{\text{Rad}}_S(L(\mathcal{H}))$, we need instead to express it using $\widehat{\text{Rad}}_S(\mathcal{H})$. Because $\widehat{\text{Rad}}_S(\mathcal{H})$ is the actual complexity of the model, not just the output loss function! Therefore, we must find a way to express $\widehat{\text{Rad}}_S(L(\mathcal{H}))$ in terms of $\widehat{\text{Rad}}_S(\mathcal{H})$.

## 2.4 express $\widehat{\text{Rad}}_S(L(\mathcal{H}))$ in terms of $\widehat{\text{Rad}}_S(\mathcal{H})$

Firstly, there is no general expression for all $\mathcal{F}$ and $\mathcal{H}$. However, imagine we have a hypothesis set $\mathcal{H}$ of binary functions, i.e., $\mathcal{H} \subset \{-1, 1\}^{\mathcal{X}}$. This notation means output of hypothesis set $\mathcal{H}$ is a sub-set of all possible $\{-1, 1\}^{\mathcal{X}}$.

Then, imagine we define a composite function where we add loss to it: $f_h \in \mathcal{F}$:

$$\mathcal{F} = L(\mathcal{H}) = \{\mathbb{1}_{\{h(x)\neq y\}} \mid h \in \mathcal{H}\} \tag{13}$$

what is then the relationship between $\widehat{\text{Rad}}_S(\mathcal{H})$ and $\widehat{\text{Rad}}_S(L(\mathcal{H}))$:

**Lemma 3** *given* $L(\mathcal{H}) = \{\mathbb{1}_{\{h(x)\neq y\}} \mid h \in \mathcal{H}\}$ *and* $h(x) \in \{-1, 1\}$*:*

$$\widehat{Rad}_S(L(\mathcal{H})) = \frac{1}{2}\widehat{Rad}_S(\mathcal{H}) \tag{14}$$

The results is as expected, as $f_h \in \mathcal{F}$ outputs $\{0, 1\}$, therefore $\widehat{\text{Rad}}_S(\mathcal{F})$ is much more restrictive than $\widehat{\text{Rad}}_S(\mathcal{H})$, hence has less complexity. The proof is easy to understand from the original source. Firstly, notice that conversion between $h \to f_h$ is:

$$\begin{aligned} f_h &= \mathbb{1}_{h(X_i)\neq Y_i} \qquad Y_i \in \{-1, 1\} \\ &= \frac{1 - Y_i h(X_i)}{2} \qquad \text{easy to see: } \frac{1 - Y_i h(X_i)}{2} = \begin{cases} 0: & Y_i = h(X_i) \\ 1: & Y_i \neq h(X_i) \end{cases} \end{aligned} \tag{15}$$

$$\begin{aligned} \widehat{\text{Rad}}_S\big(L(\mathcal{H})\big) &= \widehat{\text{Rad}}_S(\mathcal{F}) \\ &= \mathbb{E}_\sigma\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \frac{1 - Y_i h(X_i)}{2}\Big] \\ &= \mathbb{E}_\sigma\Big[\frac{1}{2n}\sum_{i=1}^{n}\sigma_i + \frac{1}{2}\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i \times (-Y_i)h(X_i)\Big] \\ &= \frac{1}{2n}\underbrace{\mathbb{E}_\sigma\Big[\sum_{i=1}^{n}\sigma_i\Big]}_{0} + \frac{1}{2}\mathbb{E}_\sigma\Big[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i \times (-Y_i)h(X_i)\Big] \\ &= \frac{1}{2}\mathbb{E}_\sigma\Big[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i h(X_i)\Big] \quad \because \Pr(\sigma_i(-Y_i)) = \Pr(\sigma_i) \text{ and } Y_i \text{ is symmetric} \\ &= \frac{1}{2}\widehat{\text{Rad}}_S(\mathcal{H}) \end{aligned} \tag{16}$$

However, the result is significant. Using lemma (3), you can then apply Theorem (1), where you can express $R(h)$ and $\hat{R}_S(h)$ in terms of the $\text{Rad}_n(\mathcal{H})$ and $\widehat{\text{Rad}}_S(\mathcal{H})$ instead of expressing them in terms of $\widehat{\text{Rad}}_S(L(\mathcal{H}))$, i.e.,:

$$\begin{aligned} &\forall h \in \mathcal{H}: \quad R(h) \leq \hat{R}_S(h) + 2\widehat{\text{Rad}}_S(L(\mathcal{H})) + 3\sqrt{\frac{\log(2/\delta)}{2n}} \\ \implies &\forall h \in \mathcal{H}: \quad R(h) \leq \hat{R}_S(h) + \widehat{\text{Rad}}_S(\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2n}} \end{aligned} \tag{17}$$

**Lemma 4** *given any function class* $\mathcal{H}$ *and constants* $a, b \in \mathbb{R}$*, if we denote a function class,* $\{f_h | f(x) = ah(x) + b\}$ *by* $a\mathcal{F} + b$*, then:*

immediately, you noticed that if one let $f_h = \frac{1 - Yh(\mathbf{x})}{2}$, then:

$$\begin{aligned} \widehat{\text{Rad}}_S(a\mathcal{H} + b) &= |a|\widehat{\text{Rad}}_S(\mathcal{H}) \\ \widehat{\text{Rad}}_S(\mathcal{F}) &= \widehat{\text{Rad}}_S\Big(\frac{1 - Y\mathcal{H}(\mathbf{x})}{2}\Big) \\ &= \Big|\frac{-Y}{2}\Big|\widehat{\text{Rad}}_S(\mathcal{H}) \\ &= \frac{1}{2}\widehat{\text{Rad}}_S(\mathcal{H}) \end{aligned} \tag{18}$$

### 2.4.1 Proof

the proof is rather straightforward:

$$\begin{aligned}
\widehat{\text{Rad}}_S(a\mathcal{H}+b) &= \mathbb{E}_\sigma\Big[\sup_{f_h\in a\mathcal{H}+b}\Big(\frac{1}{n}\sum_{i=1}^{n}\sigma_i f_h(Z_i)\Big)\Big] \\
&= \mathbb{E}_\sigma\Big[\sup_{h\in\mathcal{H}}\Big(\frac{1}{n}\sum_{i=1}^{n}\sigma_i a h(Z_i)+b\Big)\Big] \\
&= \mathbb{E}_\sigma\Big[\sup_{h\in\mathcal{H}}\Big(\frac{1}{n}\sum_{i=1}^{n}\sigma_i a h(Z_i)+\frac{1}{n}\sum_{i=1}^{n}\sigma_i b\Big)\Big] \\
&= \mathbb{E}_\sigma\Big[\sup_{h\in\mathcal{H}}\Big(\frac{1}{n}\sum_{i=1}^{n}\sigma_i a h(Z_i)\Big)\Big] \qquad \because b\mathbb{E}_\sigma\big[\sum_{i=1}^{n}\sigma_i\big]=0 \\
&= |a|\mathbb{E}_\sigma\Big[\sup_{h\in\mathcal{H}}\Big(\frac{1}{n}\sum_{i=1}^{n}\sigma_i h(Z_i)\Big)\Big] \\
&= |a|\widehat{\text{Rad}}_S(\mathcal{H})
\end{aligned} \tag{19}$$

the last two lines can be understood by:

$$\begin{aligned}
&\Pr(\sigma_i=1)=\Pr(\sigma_i=-1)=\frac{1}{2} \\
\implies &\Pr(\sigma_i a=a)=\Pr(\sigma_i a=-a)=\frac{1}{2}
\end{aligned} \tag{20}$$

# 3 Expected hypothesis bound proof

## 3.1 facts about $\sup(\cdot)$

Before we start the proof, we need to use few properties to about $\sup(\cdot)$ function. There are few properties about $\sup(\cdot)$ are all obvious, but we try to prove them formally:

1. **fact 1** $\sup_x f(x)+\sup_x g(x)\geq\sup_x\big(f(x)+g(x)\big)$

   **proof**, even though it's obvious:

$$\begin{aligned}
&\sup_{x,y\in A,x=y}\big(f(x)+g(y)\big)\leq\sup_{x,y\in A}\big(f(x)+g(y)\big) \quad \text{LHS has smaller domain set} \\
\implies &\sup_{x\in A}(f(x)+g(x))\leq\sup_{x\in A}f(x)+\sup_{x\in A}g(x)
\end{aligned} \tag{21}$$

2. **fact 2** $\sup_x f(x)-\sup_x g(x)\leq\sup_x\big(f(x)-g(x)\big)$

   could just use above and write:

$$\begin{aligned}
\sup_x\big((f(x)-g(x))+g(x)\big)&\leq\sup_x\big(f(x)-g(x)\big)+\sup_x\big(g(x)\big) \\
\implies \sup_x f(x)&\leq\sup_x\big(f(x)-g(x)\big)+\sup_x\big(g(x)\big) \\
\implies \sup_x f(x)-\sup_x g(x)&\leq\sup_x\big(f(x)-g(x)\big)
\end{aligned} \tag{22}$$

3. **fact 3**
proving direction of $\sup_x f(x) - \sup_x g(x) \leq \sup_x \big(f(x) - g(x)\big)$ only has one way. However, starting from $\sup_x \big(f(x) - g(x)\big)$ has two different routes:

$$\begin{cases} \sup_x \big(f(x) - g(x)\big) & \geq \sup_x f(x) - \sup_x g(x) \\ \sup_x \big(f(x) - g(x)\big) = \underbrace{\sup_x \big(f(x) + g(-x)\big)}_{\text{if possible for } g(x)=-g(x)} & \leq \sup_x f(x) + \sup g(-x) \end{cases} \tag{23}$$

second line is a lot more useful as we usually derive bound for LHS with a $\leq$ sign.

4. **fact 4** $\sup(\cdot)$ is a convex function
**proof** let $\bar{X} = \{x_1, \ldots, x_n\}$ and $\bar{Y} = \{y_1, \ldots, y_n\}$:

$$\theta \sup(\bar{X}) + (1-\theta) \sup \bar{Y} \geq \sup\big(\theta\bar{X} + (1-\theta)\bar{Y}\big) \tag{24}$$

to prove this, we can pick arbitrary $x_i$ and $y_k$ from the two sets respectively:

$$\begin{aligned} & \theta \sup(\bar{X}) + (1-\theta) \sup \bar{Y} \geq \theta x_i + (1-\theta) y_k \quad \forall i, k \\ \implies & \theta \sup(\bar{X}) + (1-\theta) \sup \bar{Y} \geq \sup\big(\theta\bar{X} + (1-\theta)\bar{Y}\big) \end{aligned} \tag{25}$$

## 3.2 Bound $\sup_{h\in\mathcal{H}} \left(\mathbb{E}_Z[h(Z)] - \frac{1}{n}\sum_{i=1}^n h(Z_i)\right)$ by $\mathbf{Rad}_n(\mathcal{H})$

We do not want to confuse $\mathbb{E}_Z[h(Z)]$ by $R(h)$, therefore in Eq.(9), we defined:

$$T(h) = \mathbb{E}_Z[h(Z)], \qquad \widehat{T}_S(h) = \frac{1}{n}\sum_{i=1}^n h(Z_i) \tag{26}$$

$$\begin{aligned} \phi(S) &= \sup_{h\in\mathcal{H}} \left(\mathbb{E}_Z[h(Z)] - \frac{1}{n}\sum_{i=1}^n h(Z_i)\right) \quad Z_i \in S \\ &= \sup_{h\in\mathcal{H}} \left(T(h) - \widehat{T}_S(h)\right) \end{aligned} \tag{27}$$

in word, $\phi(S)$ is maximum difference between empirical and expected hypothesis output for any $h \in \mathcal{H}$. Note that there is no absolute value here yet.
we then apply this to see the difference between two sets:

$$\begin{aligned} S &= \{Z_1, \ldots, Z_{i-1}, Z_i, Z_{i+1}, \cdots, Z_n\} \\ S' &= \{Z_1, \ldots, Z_{i-1}, Z_i', Z_{i+1}, \cdots, Z_n\} \end{aligned} \tag{28}$$

$$\begin{aligned} \phi(S) - \phi(S') &= \sup_{h\in\mathcal{H}} \Big(\underbrace{T[h] - \widehat{T}_S(h)}_{f(z)}\Big) - \sup_{h\in\mathcal{H}} \Big(\underbrace{T[h] - \widehat{T}_{S'}(h)}_{g(z)}\Big) \quad Z_i \in S, Z_i' \in S' \\ &\leq \sup_{h\in\mathcal{H}} \Big(\big(T(h) - \widehat{T}_S(h)\big) - \big(T(h) - \widehat{T}_{S'}(h)\big)\Big) \quad \text{use fact 1} \\ &= \sup_{h\in\mathcal{H}} \Big(\widehat{T}_{S'}(h) - \widehat{T}_S(h)\Big) \quad \text{remove } T(h) \text{ which is hard to deal with} \\ &= \frac{1}{n} \sup_{h\in\mathcal{H}} \Big(h(Z_i') - h(Z_i)\Big) \quad \text{use fact that only single } Z_i \neq Z_i' \text{ so, only one term remain in sum} \\ &\leq \frac{b-a}{n} \quad \text{assume a single } h \in [a, b] \end{aligned} \tag{29}$$

Equally, by symmetry, we can show $\phi(S') - \phi(S) \leq \frac{b-a}{n}$. Therefore we have:

$$\big|\phi(S') - \phi(S)\big| \leq \frac{b-a}{n} \tag{30}$$

which satisfies the condition for McDiarmid's inequality, i.e.,

$$\begin{aligned} |f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m)| &\leq c_i \\ \big|\phi(S') - \phi(S)\big| &\leq \frac{b-a}{n} \end{aligned} \tag{31}$$

the only complicated part is that $f(\cdot) \equiv \phi(\cdot)$ involves a $\sup_h$ term. We can derive:

$$\begin{aligned} \Pr\big(f(X_1, X_2, \cdots, X_n) - \mathbb{E}[f(X_1, X_2, \cdots, X_n)] \geq \epsilon\big) &\leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right) \\ \implies \Pr\big(\phi(S) - \mathbb{E}_S[\phi(S)] \geq \epsilon\big) &\leq \exp\left(-\frac{2\epsilon^2}{n\left(\frac{(b-a)}{n}\right)^2}\right) \\ &\leq \exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right) \end{aligned} \tag{32}$$

now, simplify with R.H.S:

$$\begin{aligned} \delta &= \exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right) \\ \log(\delta) &= -\frac{2n\epsilon^2}{(b-a)^2} \\ \log(1/\delta) &= \frac{2n\epsilon^2}{(b-a)^2} \\ \implies \epsilon^2 &= \frac{\log(1/\delta)(b-a)^2}{2n} \\ \implies \epsilon &= (b-a)\sqrt{\frac{\log(1/\delta)}{2n}} \end{aligned} \tag{33}$$

therefore, with probability of **at least** $1-\delta$:

$$\begin{aligned} \phi(S) - \mathbb{E}_S[\phi(S)] &\leq (b-a)\sqrt{\frac{\log(1/\delta)}{2n}} \\ \implies \phi(S) &\leq \mathbb{E}_S[\phi(S)] + (b-a)\sqrt{\frac{\log(1/\delta)}{2n}} \end{aligned} \tag{34}$$

however, the above was just bounding $\phi(S) = \sup_{h\in\mathcal{H}}\big(T(h) - \widehat{T}_S(h)\big)$, this is independent of Rademacher complexity.

### 3.2.1 introduce $\mathbf{Rad}_n(\mathcal{H})$

looking at Eq.(34), let's see if we can bound the quantity $\mathbb{E}_S[\phi(S)]$ with $\text{Rad}_n(\mathcal{H})$.

Firstly, we realize that:

$$
\begin{aligned}
T(h) &= \mathbb{E}_Z[h(Z)] \qquad \text{expectation of single sample} \\
&= \mathbb{E}_S[\widehat{T}_S(h)] \qquad \text{expectation of sample average} \\
&= \mathbb{E}_S\Big[\frac{1}{n}\sum_{i=1}^{n} h(Z_i)\Big] \qquad Z_i \in S
\end{aligned}
\tag{35}
$$

This is analogous to Empirical and Expected loss.

Note that we let $T(h) = \mathbb{E}_S\big[\frac{1}{n}\sum_{i=1}^{n} h(Z_i)\big]$ instead of $T(h) = \mathbb{E}_Z[h(Z)]$, so that the equations will align with $\widehat{\text{Rad}}_S(\mathcal{H})$.

Using the expectation of sample average version from Eq.(35). In here, we need to bring a ghost set $\tilde{S}$ into the equation, which needs to be **independent** from the $S$ in the outer expectation.

$$
\begin{aligned}
\mathbb{E}_S[\phi(S)] &= \mathbb{E}_S\Big[\sup_{h\in\mathcal{H}}\Big(T(h) - \widehat{T}_S(h)\Big)\Big] \\
&= \mathbb{E}_S\Big[\sup_{h\in\mathcal{H}}\Big(\mathbb{E}_{\tilde{S}}\Big[\frac{1}{n}\sum_{i=1}^{n} h(\tilde{Z}_i)\Big] - \frac{1}{n}\sum_{i=1}^{n} h(Z_i)\Big)\Big] \quad Z_i \in S
\end{aligned}
\tag{36}
$$

note that in here, all elements $S = \{Z_i\}$ differ from $\tilde{S} = \{\tilde{Z}_i\}$. Do not confuse this with Eq.(28), in here, **all** elements in sets $S$ and $\tilde{S}$ are independent.

$$
\begin{aligned}
\mathbb{E}_S[\phi(S)] &= \mathbb{E}_S\Big[\sup_{h\in\mathcal{H}}\Big(\mathbb{E}_{\tilde{S}}\Big[\frac{1}{n}\sum_{i=1}^{n} h(\tilde{Z}_i)\Big] - \frac{1}{n}\sum_{i=1}^{n} h(Z_i)\Big)\Big] \quad Z_i \in S, \quad \tilde{Z}_i \in \tilde{S} \\
&= \mathbb{E}_S\Big[\underbrace{\sup_{h\in\mathcal{H}}\frac{1}{n}\Big(\mathbb{E}_{\tilde{S}}\Big[\sum_{i=1}^{n}\big(h(\tilde{Z}_i) - h(Z_i)\big)\Big]\Big)}_{\sup(\mathbb{E}[\cdot])}\Big] \quad Z_i \in S, \quad \tilde{Z}_i \in \tilde{S} \\
&\le \mathbb{E}_S\Big[\underbrace{\mathbb{E}_{\tilde{S}}\Big[\sup_{h\in\mathcal{H}}\frac{1}{n}\Big(\sum_{i=1}^{n}\big(h(\tilde{Z}_i) - h(Z_i)\big)\Big)\Big]}_{\mathbb{E}[\sup(\cdot)]}\Big] \\
&= \mathbb{E}_S\Big[\mathbb{E}_{\tilde{S}}\Big[\mathbb{E}_{\bar{\sigma}}\Big[\sup_{h\in\mathcal{H}}\frac{1}{n}\Big(\sum_{i=1}^{n}\sigma_i\big(h(\tilde{Z}_i) - h(Z_i)\big)\Big)\Big]\Big]\Big] \qquad \because h(\tilde{Z}_i) - h(Z_i) \text{ are symmetric, same trick prove hoeffding lemma} \\
&= \mathbb{E}_{S,\tilde{S},\bar{\sigma}}\Big[\underbrace{\sup_{h\in\mathcal{H}}\frac{1}{n}\Big(\sum_{i=1}^{n}\sigma_i h(\tilde{Z}_i) + \big(-\sigma_i h(Z_i)\big)\Big)}_{\sup_x\left(f(x)+(-g(x))\right) \le \sup_x f(x) + \sup_x -g(x)}\Big] \quad \text{use short notation} \\
&\le \mathbb{E}_{\tilde{S},\bar{\sigma}}\Big[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i h(\tilde{Z}_i)\Big] + \mathbb{E}_{S,\bar{\sigma}}\Big[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n} -\sigma_i h(Z_i)\Big] \\
&= \mathbb{E}_{\tilde{S},\bar{\sigma}}\Big[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i h(\tilde{Z}_i)\Big] + \mathbb{E}_{S,\bar{\sigma}}\Big[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i h(Z_i)\Big] \qquad \text{both sets are from } \mathcal{D} \\
&= 2\mathbb{E}_{S,\bar{\sigma}}\Big[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i h(Z_i)\Big] \\
&= 2\text{Rad}_n(\mathcal{H})
\end{aligned}
\tag{37}
$$

therefore, with probability of **at least** $1-\delta$:

$$\begin{aligned} \phi(S) &\le \mathbb{E}_S[\phi(S)] + (b-a)\sqrt{\frac{\log(1/\delta)}{2n}} \\ \sup_{h\in\mathcal{H}}\Big(T(h) - \widehat{T}_S(h)\Big) &\le 2\text{Rad}_n(\mathcal{H}) + (b-a)\sqrt{\frac{\log(1/\delta)}{2n}} \end{aligned} \tag{38}$$

**Fatou lemma**

$$\underbrace{\sup_{h\in\mathcal{H}} \frac{1}{n}\Big(\mathbb{E}_{S'}\Big[\sum_{i=1}^{n}\big(h(Z_i') - h(Z_i)\big)\Big]\Big)}_{\sup(\mathbb{E}[\cdot])} \le \underbrace{\mathbb{E}_{S'}\Big[\sup_{h\in\mathcal{H}}\frac{1}{n}\Big(\sum_{i=1}^{n}\big(h(Z_i') - h(Z_i)\big)\Big)\Big]}_{\mathbb{E}[\sup(\cdot)]} \tag{39}$$

Fatou lemmas says:

$$\limsup_{n\to\infty}\int_S f_n\, d\mu \le \int_S \limsup_{n\to\infty} f_n\, d\mu. \tag{40}$$

however, as we haven't chosen $n\to\infty$, it is unclear how Fatou Lemma is applied.

## 3.3 bound $\text{Rad}_n(\mathcal{H})$ using $\widehat{\text{Rad}}_S(\mathcal{H})$

The result in Eq.(38) are in terms of $\text{Rad}_n(\mathcal{H})$, which may be difficult to evaluate.

We hope to apply this to $\widehat{\text{Rad}}_S(\mathcal{H})$ making it possible to evaluate. Similar to before, We can apply McDiarmid's inequality. To do so, we introduce a set which differ by a single element $S'$:

$$\begin{aligned} S &= \{Z_1, \dots, Z_{i-1}, Z_i, Z_{i+1}, \cdots, Z_n\} \\ S' &= \{Z_1, \dots, Z_{i-1}, Z_i', Z_{i+1}, \cdots, Z_n\} \end{aligned} \tag{41}$$

then,

$$\begin{aligned} \widehat{\text{Rad}}_S(\mathcal{H}) - \widehat{\text{Rad}}_{S'}(\mathcal{H}) &= \mathbb{E}_{\bar\sigma}\Big[\sup_{h\in\mathcal{H}}\frac{\sum_{Z_i\in S}\sigma_i h(Z_i)}{n}\Big] - \mathbb{E}_{\bar\sigma}\Big[\sup_{h\in\mathcal{H}}\frac{\sum_{Z_i\in S'}\sigma_i h(Z_i)}{n}\Big] \\ &= \mathbb{E}_{\bar\sigma}\Big[\sup_{h\in\mathcal{H}}\frac{\sum_{Z_i\in S}\sigma_i h(Z_i)}{n} - \sup_{h\in\mathcal{H}}\frac{\sum_{Z_i\in S'}\sigma_i h(Z_i)}{n}\Big] \\ &= \frac{1}{n}\mathbb{E}_{\bar\sigma}\Big[\sup_{h\in\mathcal{H}}\sum_{Z_i\in S}\sigma_i h(Z_i) - \sup_{h\in\mathcal{H}}\sum_{Z_i\in S'}\sigma_i h(Z_i)\Big] \\ &\le \frac{1}{n}\mathbb{E}_{\bar\sigma}\Big[\sup_{h\in\mathcal{H}}\Big(\sum_{Z_i\in S}\sigma_i h(Z_i) - \sum_{Z_i\in S'}\sigma_i h(Z_i)\Big)\Big] \\ &= \frac{1}{n}\mathbb{E}_{\sigma_i}\Big[\sup_{h_i\in\mathcal{H}}\sigma_i\big(h(Z_i) - \sigma_i h(Z_i')\big)\Big] \\ &= \frac{1}{n}\sup_{h_i\in\mathcal{H}}\big(h(Z_i) - h(Z_i')\big) \qquad \text{symmetry} \\ &\le \frac{b-a}{n} \end{aligned} \tag{42}$$

then, by proving $\widehat{\text{Rad}}_{S'}(\mathcal{H}) - \widehat{\text{Rad}}_S(\mathcal{H}) \le \frac{b-a}{n}$, let to the condition (not probability) for McDiarmid:

$$\left|\widehat{\text{Rad}}_S(\mathcal{H}) - \widehat{\text{Rad}}_{S'}(\mathcal{H})\right| \leq \frac{b-a}{n} \tag{43}$$

Apply McDiarmid's inequality, we can bound the function by the expectation (we need it for $\text{Rad}_n(\mathcal{H})$!):

$$\begin{aligned}
\widehat{\text{Rad}}_S(\mathcal{H}) - \mathbb{E}_S\big[\text{Rad}_S(\mathcal{H})\big] &= \widehat{\text{Rad}}_S(\mathcal{H}) - \text{Rad}_n(\mathcal{H}) \\
&\leq \exp\Big(-\frac{2n\epsilon^2}{(b-a)^2}\Big) \\
\implies \Pr\Big(\text{Rad}_n(\mathcal{H}) \geq \widehat{\text{Rad}}_S(\mathcal{H}) + (b-a)\sqrt{\frac{\log(1/\delta)}{2n}}\Big) &\leq \delta \\
\implies \Pr\Big(\text{Rad}_n(\mathcal{H}) \leq \widehat{\text{Rad}}_S(\mathcal{H}) + (b-a)\sqrt{\frac{\log(1/\delta)}{2n}}\Big) &\geq 1-\delta
\end{aligned} \tag{44}$$

## 3.4 combine RHS using both $\text{Rad}_n(\mathcal{H})$ and $\widehat{\text{Rad}}_S(\mathcal{H})$

$$\begin{aligned}
&\Pr\Big(\text{Rad}_n(\mathcal{H}) \geq \widehat{\text{Rad}}_S(\mathcal{H}) + (b-a)\sqrt{\frac{\log(1/\delta)}{2n}}\Big) \leq \delta \\
\implies &\Pr\Big(\text{Rad}_n(\mathcal{H}) \geq \widehat{\text{Rad}}_S(\mathcal{H}) + (b-a)\sqrt{\frac{\log(2/\delta)}{2n}}\Big) \leq \frac{\delta}{2} \\
\implies &\Pr\Big(2\text{Rad}_n(\mathcal{H}) \geq \underbrace{2\widehat{\text{Rad}}_S(\mathcal{H}) + 2(b-a)\sqrt{\frac{\log(2/\delta)}{2n}}}\Big) \leq \frac{\delta}{2}
\end{aligned} \tag{45}$$

From **theorem** (1):

$$\Pr\Big(T(h) \geq \widehat{T}_S(h) + 2\text{Rad}_n(\mathcal{H}) + (b-a)\sqrt{\frac{\log(2/\delta)}{2n}}\Big) \leq \frac{\delta}{2} \tag{46}$$

substitute Eq.(45) into Theorem (1):

$$\begin{aligned}
\implies \Pr\Big(T(h) \geq \widehat{T}_S(h) + \underbrace{2\widehat{\text{Rad}}_S(\mathcal{H}) + 2(b-a)\sqrt{\frac{\log(2/\delta)}{2n}}} + (b-a)\sqrt{\frac{\log(2/\delta)}{2n}}\Big) &\leq \frac{\delta}{2} + \frac{\delta}{2} \qquad \text{union bound} \\
\implies \Pr\Big(T(h) \geq \widehat{T}_S(h) + 2\widehat{\text{Rad}}_S(\mathcal{H}) + 3(b-a)\sqrt{\frac{\log(2/\delta)}{2n}}\Big) &\leq \delta
\end{aligned} \tag{47}$$

# 4 Finite Class Lemma (Massart)

**Lemma 5** *Let $\mathcal{A}$ be some finite subset of $\mathbb{R}^m$ and let:*

$$r = \sup_{\mathbf{a}\in\mathcal{A}} \|\mathbf{a}\| \quad or \quad \|\mathbf{a}\| \leq r \tag{49}$$

*then:*

$$\mathbb{E}_{\bar{\sigma}}\Big[\sup_{\mathbf{a}\in\mathcal{A}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i a_i\Big] \leq \frac{r\sqrt{2\log(|\mathcal{A}|)}}{n} \tag{50}$$

It is obvious that we are using it for Rademarcher complexity. To make it look like $\mathbb{E}_{\bar{\sigma}}\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i h(Z_i)\Big]$, by equating:

$$\begin{aligned} a_i &\equiv h(Z_i) \\ \mathbf{a} &= [h(Z_1) \quad h(Z_2) \quad \dots \quad h(Z_n)] \end{aligned} \tag{51}$$

Also $|\mathcal{A}|$ is bounded, i.e., we have a finite class, meaning the value of $h(Z)$ is finite.

## 4.1 proof

we can just prove **lemma 5** using $\mathbb{E}_{\bar{\sigma}}\Big[\sup_{a\in\mathcal{A}}\sum_{i=1}^{n}\sigma_i a_i\Big] \leq r\sqrt{2\log(|A|)}$:

$$
\begin{aligned}
\exp\Big(\lambda\mathbb{E}_{\bar\sigma}\Big[\sup_{\mathbf{a}\in\mathcal{A}}\sum_{i=1}^{n}\sigma_i a_i\Big]\Big) &\le \mathbb{E}_{\bar\sigma}\Big[\exp\Big(\lambda\sup_{\mathbf{a}\in\mathcal{A}}\sum_{i=1}^{n}\sigma_i a_i\Big)\Big] \quad \lambda>0: \quad \text{Jensen}\\
&= \mathbb{E}_{\bar\sigma}\Big[\sup_{\mathbf{a}\in\mathcal{A}}\Big(\exp\Big(\lambda\sum_{i=1}^{n}\sigma_i a_i\Big)\Big)\Big] \quad \exp(\lambda(\cdot)) \text{ monotone}\\
&\le \mathbb{E}_{\bar\sigma}\Big[\sum_{\mathbf{a}\in\mathcal{A}}\exp\Big(\lambda\sum_{i=1}^{n}\sigma_i a_i\Big)\Big] \quad \text{replace sup with } \sum\\
&= \sum_{\mathbf{a}\in\mathcal{A}}\mathbb{E}_{\bar\sigma}\Big[\exp\Big(\lambda\sum_{i=1}^{n}\sigma_i a_i\Big)\Big]\\
&= \sum_{\mathbf{a}\in\mathcal{A}}\mathbb{E}_{\bar\sigma}\Big[\prod_{i=1}^{n}\exp^{\lambda\sigma_i a_i}\Big]\\
&= \sum_{\mathbf{a}\in\mathcal{A}}\prod_{i=1}^{n}\mathbb{E}_{\sigma_i}\big[\exp^{\lambda\sigma_i a_i}\big] \quad \because \Pr(\bar\sigma)=\prod_{i}^{n}\Pr(\sigma_i)\\
&= \sum_{\mathbf{a}\in\mathcal{A}}\prod_{i=1}^{n}\Big(\frac{1}{2}\exp^{\lambda(+1)a_i}+\frac{1}{2}\big[\exp^{\lambda(-1)a_i}\big]\Big) \quad \text{write expectation explicitly}\\
&= \sum_{\mathbf{a}\in\mathcal{A}}\prod_{i=1}^{n}\frac{\exp^{\lambda a_i}+\exp^{-\lambda a_i}}{2}\\
&\le \sum_{\mathbf{a}\in\mathcal{A}}\prod_{i=1}^{n}\exp\Big(\frac{\lambda^2 a_i^2}{2}\Big) \quad \because \frac{\exp^{x}+\exp^{-x}}{2}\le\exp^{\frac{x^2}{2}}\\
&\qquad\qquad \text{remember previous class: } \underbrace{\frac{\exp^{\lambda}+\exp^{-\lambda}}{2}}_{\text{MGF}_{\sigma\sim Rad(\lambda)}} \le \underbrace{\exp^{\frac{\lambda^2}{2}}}_{\text{Hoeffding lemma bound}}\\
&= \sum_{\mathbf{a}\in\mathcal{A}}\exp\Big(\sum_{i=1}^{n}\frac{\lambda^2 a_i^2}{2}\Big)=\sum_{\mathbf{a}\in\mathcal{A}}\exp\Big(\frac{\lambda^2}{2}\sum_{i=1}^{n}a_i^2\Big)\\
&= \sum_{\mathbf{a}\in\mathcal{A}}\exp\Big(\frac{\lambda^2}{2}\|\mathbf{a}\|^2\Big)\\
&\le |\mathcal{A}|\exp^{\frac{\lambda^2 r^2}{2}} \quad \text{union bound over } \textbf{finite} \text{ class set } \mathcal{A}
\end{aligned}
\tag{52}
$$

let $\mu=\mathbb{E}_{\bar\sigma}\Big[\sup_{\mathbf{a}\in\mathcal{A}}\sum_{i=1}^{n}\sigma_i a_i\Big]$:

$$
\begin{aligned}
\exp\Big(\lambda\mathbb{E}_{\bar\sigma}\Big[\sup_{\mathbf{a}\in\mathcal{A}}\sum_{i=1}^{n}\sigma_i a_i\Big]\Big)\equiv\exp\big(\lambda\mu\big) &\le |\mathcal{A}|\exp^{\frac{\lambda^2 r^2}{2}}\\
\lambda\mu &\le \log(|\mathcal{A}|)+\frac{\lambda^2 r^2}{2}\\
\implies \mu &\le \frac{\log(|\mathcal{A}|)}{\lambda}+\frac{\lambda r^2}{2}
\end{aligned}
\tag{53}
$$

$$
\begin{aligned}
-\log(|\mathcal{A}|)\lambda^{-2} + \frac{r^2}{2} &= 0 \\
\implies \frac{r^2}{2\log(|\mathcal{A}|)} &= \lambda^{-2} \\
\implies \lambda^2 &= \frac{2\log(|\mathcal{A}|)}{r^2} \\
\implies \lambda &= \frac{\sqrt{2\log(|\mathcal{A}|)}}{r}
\end{aligned}
\tag{54}
$$

by substitution, we have:

$$
\begin{aligned}
\mu &\leq \log(|\mathcal{A}|)\frac{r}{\sqrt{2\log(|\mathcal{A}|)}} + \frac{\sqrt{2\log(|\mathcal{A}|)}r}{2} \\
&= \frac{\sqrt{2\log(|\mathcal{A}|)}r}{2} + \frac{\sqrt{2\log(|\mathcal{A}|)}r}{2} \\
&= r\sqrt{2\log(|\mathcal{A}|)}
\end{aligned}
\tag{55}
$$

**Corollary 5.1** *Let $\mathcal{H}$ be finite set of functions such that $|h(z)| \leq 1$:*

$$
\mathit{Rad}_n(\mathcal{H}) \leq \sqrt{\frac{2\log|\mathcal{H}|}{n}} \tag{56}
$$

Given $\mathcal{H}$ and $S = \{z_1, \ldots, z_n\}$ and let $\big\{\mathbf{h} \equiv \big(h(z_1), \ldots, h(z_n)\big) : h \in \mathcal{H}\big\}$, then for every $h(z) \in \mathcal{H}$, we have:

$$
\|\mathbf{a}\| \leq \sqrt{\sum_{i=1}^{n} h(z_i)} \leq \sqrt{n} \tag{57}
$$

apply **Massart Lemma**, we have $r \equiv \sqrt{n}$ and $a_i \equiv h(z_i)$:

$$
\begin{aligned}
\mathbb{E}_{\bar{\sigma}}\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i h(z_i)\Big] &\leq \sqrt{n}\frac{\sqrt{2\log(|\mathcal{H}|)}}{n} \\
&= \sqrt{\frac{2\log(|\mathcal{H}|)}{n}}
\end{aligned}
\tag{58}
$$

# 5 Rademacher Calculus

## 5.1 $c\cdot\mathcal{H}+b = \{c\cdot h+b : h \in \mathcal{H}\} \implies \mathbf{Rad}_n(c\cdot\mathcal{H}+b) = |c|\mathbf{Rad}_n(\mathcal{H})$

this is already proved in lemma 4.

## 5.2 $\mathbf{Rad}_n(\mathbf{conv}(\mathcal{H})) = \mathbf{Rad}_n(\mathcal{H})$

**Lemma 6** *Let $\mathit{conv}\mathcal{H} = \big\{\sum\theta_i h_i : \{h_i\} \subseteq \mathcal{H}, \quad \theta_i \geq 0, \sum\theta_i = 1\big\}$*
*Then:*

$$
\mathit{Rad}_n\big(\mathit{conv}(\mathcal{H})\big) = \mathit{Rad}_n(\mathcal{H}) \tag{59}
$$

Rademacher complexity of all convex combination of functions does not change the original Rademacher complexity value. This is not so surprising, as convex combination does not change the overall complexity.

### 5.2.1 proof

$$
\begin{aligned}
\text{Rad}_n(\text{conv}(\mathcal{H})) &= \mathbb{E}\Big[\sup_{h_j\in\mathcal{H},\|\boldsymbol{\theta}\|_1=1}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\sum_{j=1}^{k}\theta_j h_j(X_i)\Big] \\
&= \mathbb{E}\Big[\sup_{h_j\in\mathcal{H},\|\boldsymbol{\theta}\|_1=1}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\sum_{j=1}^{k}\theta_j h_j(X_i)\Big] \\
&= \mathbb{E}\Big[\sup_{h_j\in\mathcal{H}}\sup_{\|\boldsymbol{\theta}\|_1=1}\frac{1}{n}\sum_{j=1}^{k}\theta_j\Big(\sum_{i=1}^{n}\sigma_i h_j(X_i)\Big)\Big] \\
&= \mathbb{E}\Big[\sup_{h_j\in\mathcal{H}}\max_j\frac{1}{n}\sum_{i=1}^{n}\sigma_i h_j(X_i)\Big] \qquad \text{given } \{h_j\} \text{ in outter } \sup, \text{ the inner } \boldsymbol{\theta} \text{ becomes one hot} \\
&= \mathbb{E}\Big[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i h(X_i)\Big] \\
&= \text{Rad}_n(\mathcal{H})
\end{aligned}
\tag{60}
$$

meant learning over distributions of hypotheses is statistically not any harder than learning over individual hypotheses.

### 5.2.2 application of $\textbf{Rad}_n\big(\textbf{conv}(\mathcal{H})\big) = \textbf{Rad}_n(\mathcal{H})$

**Lemma 7**

$$
\textit{conv}(S_{n,s}) \subset K_{n,s} \subset 2\textit{conv}(S_{n,s}) \tag{61}
$$

*where conv$(\cdot)$ is the convex hull that contains the set $S_{n,s}$*

$$
\begin{aligned}
K_{n,s} &:= \{\mathbf{x}\in\mathbb{R}^n : \|\mathbf{x}\|_2 \le 1, \|\mathbf{x}\|_1 \le \sqrt{s}\} = B_2^n \cap \sqrt{s}B_1^n\} \\
S_{n,s} &:= \{\mathbf{x}\in\mathbb{R}^n : \|\mathbf{x}\|_1 \le 1, \|\mathbf{x}\|_0 \le s\}
\end{aligned}
\tag{62}
$$

1. conv$(S_{n,s}) \subset K_{n,s}$

   $S_{n,s}$ is **not** a convex set, but $K_{n,s}$ is the convex set. We need to prove that, $\forall\mathbf{x}\in S_{n,s} \implies \mathbf{x}\in K_{n,s}$, i.e.,

$$
\|\mathbf{x}\|_1 \le 1 \cap \|\mathbf{x}\|_0 \le s \quad\implies\quad \|\mathbf{x}\|_2 \le 1 \cap \|\mathbf{x}\|_0 \le \sqrt{s} \tag{63}
$$

   firstly, $\|\mathbf{x}\| \le 1 \implies \|\mathbf{x}\|_2 \le 1$. think about the geometric region. Then:

$$
\begin{aligned}
\|\mathbf{x}\|_1 = |\mathbf{x}|^\top\mathbf{1}_s \qquad & \text{where } |\mathbf{x}| = [|x_1| \quad |x_2| \quad \dots \quad |x_n|] \\
& \mathbf{1}_s \text{ is a vector with at most } s \text{ one, and rest are zero} \\
\le \|\mathbf{x}\|_2\,\|\mathbf{1}_s\|_2 \qquad & \text{cauchy schwarz inequality} \\
\le \|\mathbf{x}\|_2\sqrt{s} \qquad & \because \|\mathbf{x}\|_0 \le s \\
\le \sqrt{s} \qquad & \because \|\mathbf{x}\|_1 \le 1 \implies \|\mathbf{x}\|_2 \le 1
\end{aligned}
\tag{64}
$$

   The convex hull conv$(S)$ is the **smallest** convex set containing $S$. Formally, the convex hull of the set $S$ of points in $n$ dimension is the intersection of all convex sets containing $S$. For $N$ points $\mathbf{x}_1,\dots,\mathbf{x}_N$, the convex hull $C$ is:

$$C = \Big\{ \sum_{j=1}^{N} \lambda_j \mathbf{x}_j : \qquad \lambda_j >= 0 \quad \forall j \quad \sum_{j=1}^{N} \lambda_j = 1 \Big\} \tag{65}$$

since $S_{n,s}$ is the convex hull and since $K_{n,s}$ is a convex set, then it must implie:

$$\text{conv}(S_{n,s}) \subset K_{n,s} \tag{66}$$

2. $K_{n,s} \subset 2\text{conv}(S_{n,s})$
   will prove it for the next session

here comes the real example:
In sparse linear regression, we want to find a linear predictor with at most s-nonzero coordinates. Loss function is:

$$\arg\min_{w} \frac{1}{2n} \sum_{i=1}^{n} (\langle w, x_i \rangle - y_i)^2 \qquad \text{s.t. } \textbf{function class} \tag{67}$$

where **function class** can be defined in terms of a ball:

1. $\mathcal{W}_0$:

$$\begin{aligned} \mathcal{W}_0 &= \{w \in \mathbb{R}^d : \|w\|_2^2 \leq 1, \|w\|_0 \leq s\} \\ \arg\min_{w} \frac{1}{2n} \sum_{i=1}^{n} (\langle w, x_i \rangle - y_i)^2 &\qquad \text{s.t. } w \in \mathcal{W}_0 \end{aligned} \tag{68}$$

where $\mathcal{H}_0 = \{f_w(x) = \langle w, x \rangle, w \in \mathcal{W}_0\}$ is NP-hard.

2. $\mathcal{W}_1$:
   However, if we work over $L_1$ ball:

$$\begin{aligned} \mathcal{W}_1 &= \{w \in \mathbb{R}^d : \quad \|w\|_2^2 \leq 1, \|w\|_1 \leq \sqrt{s}\} \\ \arg\min_{w} \frac{1}{2n} \sum_{i=1}^{n} (\langle w, x_i \rangle - y_i)^2 &\qquad \text{s.t. } w \in \mathcal{W}_1 \end{aligned} \tag{69}$$

where $\mathcal{H}_1 = \{f_w(x) = \langle w, x \rangle >, w \in \mathcal{W}_1\}$:
then optimization problem is convex, and by applying lemma 7, then we have:

$$\widehat{Rad}_n(\mathcal{W}_1) \leq 2\widehat{Rad}_n(\mathcal{W}_0) \tag{70}$$

Statistically, working with $L_1$ ball does not substantially increase the complexity.

## 5.3 $L$-Lipschitz functions

**Lemma 8** *Let $\phi_z$ be L-Lipschitz functions for every $z \in \mathcal{X}$, i.e. $|\phi_z(a) - \phi_z(b)| \leq L|a-b|$), or in our context, $|\phi(f(z)) - \phi(f'(z))| \leq L|f(z) - f(z')|$). Denote:*

$$\phi \circ \mathcal{H} = \{\phi_z(h(z)) : h \in \mathcal{H}\} \tag{71}$$

*Then:*

$$\mathit{Rad}_n(\phi \circ \mathcal{H}) = L\ \mathit{Rad}_n(\mathcal{H}) \tag{72}$$

it can be written as to proving:

$$\mathbb{E}_{\bar{\sigma}}\Big[\sup_{h\in\mathcal{H}} \frac{\sum_{i=1}^n \sigma_i \phi(h(Z_i))}{n}\Big] \leq L\mathbb{E}_{\bar{\sigma}}\Big[\sup_{h\in\mathcal{H}} \frac{\sum_{i=1}^n \sigma_i h(Z_i)}{n}\Big] \tag{73}$$

let's remove $n$ on both sides, and write out expectation of $h(Z_i)$ explicitly and w.l.o.g, we work with just one element $\sigma_1$, which can then be replicated to other elements throughout:

$$\begin{aligned}
&\mathbb{E}_{\bar{\sigma}}\Big[\sup_{h\in\mathcal{H}} \Big(\sigma_1\phi(h(Z_1)) + \sum_{i=2}^n \sigma_i\phi(h(Z_i))\Big)\Big] \\
&= \mathbb{E}_{\sigma_1}\Big[\mathbb{E}_{\sigma_2,\dots\sigma_n}\Big[\sup_{h\in\mathcal{H}} \Big(\sigma_1\phi(h(Z_1)) + \sum_{i=2}^n \sigma_i\phi(h(Z_i))\Big)\Big]\Big] \\
&= \underbrace{\frac{1}{2}}_{\Pr(\sigma_1=+1)}\Big[\mathbb{E}_{\sigma_2,\dots\sigma_n}\Big[\sup_{h\in\mathcal{H}} \Big(\underbrace{(+1)}_{\sigma_1=1}\phi(h(Z_1)) + \sum_{i=2}^n \sigma_i\phi(h(Z_i))\Big)\Big]\Big] + \underbrace{\frac{1}{2}}_{\Pr(\sigma_1=-1)}\Big[\mathbb{E}_{\sigma_2,\dots\sigma_n}\Big[\sup_{h\in\mathcal{H}} \Big(\underbrace{(-1)}_{\sigma_1=-1}\phi(h(Z_1)) + \sum_{i=2}^n \sigma_i\phi(h(Z_i))\Big)\Big]\Big] \\
&= \frac{1}{2}\Big[\mathbb{E}_{\sigma_2,\dots\sigma_n}\Big[\sup_{h\in\mathcal{H}} \Big(\phi(h(Z_1)) + \sum_{i=2}^n \sigma_i\phi(h(Z_i))\Big)\Big]\Big] + \frac{1}{2}\Big[\mathbb{E}_{\sigma_2,\dots\sigma_n}\Big[\sup_{h\in\mathcal{H}} \Big(-\phi(h(Z_1)) + \sum_{i=2}^n \sigma_i\phi(h(Z_i))\Big)\Big]\Big] \\
&= \frac{1}{2}\Big(\mathbb{E}_{\sigma_2,\dots\sigma_n}\Big[\sup_{h\in\mathcal{H}} \Big(\phi(h(Z_1)) + \sum_{i=2}^n \sigma_i\phi(h(Z_i))\Big) + \sup_{h\in\mathcal{H}} \Big(-\phi(h(Z_1)) + \sum_{i=2}^n \sigma_i\phi(h(Z_i))\Big)\Big]\Big) \\
&= \frac{1}{2}\mathbb{E}_{\sigma_2,\dots\sigma_n}\Big[\sup_{h,h'\in\mathcal{H}} \Big(\phi(h(Z_1)) + \sum_{i=2}^n \sigma_i\phi(h(Z_i)) - \phi(h'(Z_1)) + \sum_{i=2}^n \sigma_i\phi(h'(Z_i))\Big)\Big] \quad \text{still two separate } \sup(\cdot) \\
&\leq \frac{1}{2}\mathbb{E}_{\sigma_2,\dots\sigma_n}\Big[\sup_{h,h'\in\mathcal{H}} \Big(L\big|h(Z_1) - h'(Z_1)\big| + \sum_{i=2}^n \sigma_i\phi(h(Z_i)) + \sum_{i=2}^n \sigma_i\phi(h'(Z_i))\Big)\Big] \quad \because |\phi(h(z)) - \phi(h'(z))| \leq L|h(z) - h'(z)|) \\
&= \frac{1}{2}\mathbb{E}_{\sigma_2,\dots\sigma_n}\Big[\sup_{h,h'\in\mathcal{H}} \Big(L(h(Z_1) - h'(Z_1)) + \sum_{i=2}^n \sigma_i\phi(h(Z_i)) + \sum_{i=2}^n \sigma_i\phi(h'(Z_i))\Big)\Big]
\end{aligned} \tag{74}$$

the reason you can remove absolute operator in $\big|h(Z_1) - h'(Z_1)\big| \to (h(Z_1) - h'(Z_1))$, because if we swap $h$ with $h'$ it won't affect the sum $\sum_{i=2}^n \sigma_i h(Z_i) + \sum_{i=2}^n \sigma_i h'(Z_i)$, so sup can swap $h$ with $h'$ to make $h(Z_1) - h'(Z_1) \geq 0$

$$
\begin{aligned}
&\mathbb{E}_{\bar{\sigma}}\Big[\sup_{h\in\mathcal{H}}\Big(\sigma_1\phi(h(Z_1))+\sum_{i=2}^{n}\sigma_i\phi(h(Z_i))\Big)\Big]\\
&=\frac{1}{2}\mathbb{E}_{\sigma_2,\ldots\sigma_n}\Big[\sup_{h\in\mathcal{H}}\Big(Lh(Z_1)+\sum_{i=2}^{n}\sigma_i\phi(h(Z_i))\Big)+\sup_{\sup_{h\in\mathcal{H}}}\Big(-Lh(Z_1)+\sum_{i=2}^{n}\sigma_i\phi(h(Z_i))\Big)\Big]\\
&=\underbrace{\frac{1}{2}}_{\Pr(\sigma_1=1)}\mathbb{E}_{\sigma_2,\ldots\sigma_n}\Big[\sup_{h\in\mathcal{H}}\Big(\underbrace{+1}_{\sigma_1=1}Lh(Z_1)+\sum_{i=2}^{n}\sigma_i\phi(h(Z_i))\Big)\Big]+\underbrace{\frac{1}{2}}_{\Pr(\sigma_1=-1)}\mathbb{E}_{\sigma_2,\ldots\sigma_n}\Big[\sup_{h\in\mathcal{H}}\Big(\underbrace{-1}_{\sigma_1=-1}Lh(Z_1)+\sum_{i=2}^{n}\sigma_i\phi(h(Z_i))\Big)\Big]\\
&=\mathbb{E}_{\sigma_1,\ldots\sigma_n}\Big[\sup_{h\in\mathcal{H}}\Big(L\sigma_1h(Z_1)+\sum_{i=1}^{n}\sigma_i\phi(h(Z_i))\Big)\Big]
\end{aligned}
\tag{75}
$$

we can show the rest by computing the expectation using just $\sigma_2$ for the following:

$$
\begin{aligned}
&\mathbb{E}_{\bar{\sigma}}\Big[\sup_{h\in\mathcal{H}}\Big(L\sigma_1h(Z_1)+\sigma_2\phi(h(Z_2))+\sum_{i=3}^{n}\sigma_i\phi(h(Z_i))\Big)\Big]\\
&=\mathbb{E}_{\sigma_2}\Big[\mathbb{E}_{\sigma_1,\sigma_3,\ldots\sigma_n}\Big[\sup_{h\in\mathcal{H}}\Big(L\sigma_1h(Z_1)+\sigma_2\phi(h(Z_2))+\sum_{i=3}^{n}\sigma_i\phi(h(Z_i))\Big)\Big]
\end{aligned}
\tag{76}
$$

repeat the process we will obtain:

$$
\mathbb{E}_{\bar{\sigma}}\Big[\sup_{h\in\mathcal{H}}\frac{\sum_{i=1}^{n}\sigma_i\phi(h(Z_i))}{n}\Big]\leq L\mathbb{E}_{\bar{\sigma}}\Big[\sup_{h\in\mathcal{H}}\frac{\sum_{i=1}^{n}\sigma_ih(Z_i)}{n}\Big]
\tag{77}
$$