Machine Learning Theory Lecture 3: Rademacher Complexity

Richard Xu

October 12, 2021

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) finite $h(Z_i)$ outcomes,

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

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

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

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

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

1.2.1 can help to bound expected function value:

For example:

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

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

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

1.2.2 it can also help to bound expected risk

Theorem 2 *let $\mathcal{H}$ be set of hypothesis taking values in $\{-1, +1\}$ and for any $\delta > 0$, with probability at least $1 - \delta$ over a sample S of size n drawn from $\mathcal{D}$:*

$$\forall h \in \mathcal{H} : \quad R(h) \leq \hat{R}_S(h) + \widehat{Rad}_S(\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2n}} \tag{7}$$

2 Rademacher Complexity: generic binary functions

First up, let's not worry about the type of $\mathcal{H}$ we use, we only know it's a generic binary function:

Theorem 3 *let $\mathcal{H}$ be a set of binary functions. Then, for all n:*

$$Rad_n(\mathcal{H}) \leq \sqrt{\frac{2\log s(\mathcal{H}, n)}{n}} \tag{8}$$

2.1 proof

2.1.1 change where $\max$ is over

obviously, trying to max over $\mathcal{H}$ in $\sup_{h\in\mathcal{H}} \frac{\sum_{i=1}^n \sigma_i h(Z_i)}{n}$ is difficult, as $|\mathcal{H}|$ can be infinite . Luckily, the output $\mathcal{H}_{Z_1,\ldots,Z_n}$ is finite:

$\mathcal{H}_{Z_1,\ldots,Z_n}$ maps a particular input $Z_1, \ldots, Z_n$ into a set of binary values (by trying out all $h \in \mathcal{H}$). For example $n = 4$:

$$\mathbf{V}_n = \mathcal{H}_{Z_1,\ldots,Z_n} = \{ \underbrace{(0,0,0,1)}_{V_1}, \underbrace{(0,0,1,1)}_{V_2}, \ldots, \underbrace{(0,0,1,1)}_{V_{|\mathbf{V}_n|}} \} \tag{9}$$

of course, there must be a particular $\bar{Z}$ gives most number of different output. Therefore **shattering number** is:

$$\begin{aligned} s(\mathcal{H}, n) &= \max_{\bar{Z}} |\mathcal{H}_{Z_1,\ldots,Z_n}| \\ &= \max_{\bar{Z}} |\mathbf{V}_n| \leq 2^n \end{aligned} \tag{10}$$

$$\begin{aligned} \text{Rad}_n(\mathcal{H}) &= \mathbb{E}_{\bar{Z}}\Big[\mathbb{E}_{\bar{\sigma}}\Big[\sup_{h\in\mathcal{H}} \frac{\sum_{i=1}^n \sigma_i h(Z_i)}{n}\Big]\Big] \\ &\text{rewrite the max over set: } \mathcal{H} \to \mathcal{H}_{Z_1,\ldots,Z_n} : \\ &= \mathbb{E}_{\bar{Z}}\Big[\mathbb{E}_{\bar{\sigma}}\Big[\max_{V_j \in \mathcal{H}_{Z_1,\ldots,Z_n}} \Big\{\frac{\sum_{i=1}^n \sigma_i V_{j,i}}{n}\Big\}_{j=1}^{|\mathbf{V}_n|} \Bigm| \bar{Z}\Big]\Big] \\ &= \mathbb{E}_{\bar{Z}}\Big[\mathbb{E}_{\bar{\sigma}}\Big[\max_{V_j \in \mathcal{H}_{Z_1,\ldots,Z_n}} \Big\{\frac{\bar{\sigma}^\top V_j}{n}\Big\}_{j=1}^{|\mathbf{V}_n|} \Bigm| \bar{Z}\Big]\Big] \end{aligned} \tag{11}$$

2.1.2 Bound the inner term

Note that $V_{j,i} = h(Z_i)$, and obviously is a random variable since Z_i is random, but as for the inner expectation is concerned, it is fixed.

$$\mathbb{E}_{\bar{\sigma}}\Big[\max_{V_j \in \mathcal{H}_{Z_1,\ldots,Z_n}} \Big\{\frac{\bar{\sigma}^\top V_j}{n}\Big\}_{j=1}^{|\mathbf{V}_n|} \Bigm| \bar{Z}\Big] \tag{12}$$

where V_j is treated as constant, and since we have an expectation of maximum, we can use using **Theorem(5)**, i..e, $\mathbb{E}\big[\max\{X_1,\ldots,X_n\}\big] \leq t\sqrt{2\log(n)}$. But before we can use **Theorem(5)**, we need to show $\Big(\sum_{i=1}^n \frac{\sigma_i v_{j,i}}{n}\Big) \sim \text{SubG}(\frac{1}{n})$

2.1.3 what is $\mathbb{E}\left[\sum_{i=1}^n \frac{\sigma_i v_{j,i}}{n}\right]$?

since $\frac{\sigma_i v_i}{n}$ has zero mean, therefore, the sum also has zero mean.

2.1.4 show $\Big(\sum_{i=1}^n \frac{\sigma_i v_{j,i}}{n}\Big) \sim$ SubG$(\frac{1}{n})$

From Lecture 2, in Eq.(**??**), we know:

$$\mathbb{E}_{\sigma\sim\text{Rad}}[\exp^{\lambda\sigma}] \leq \exp\left(\frac{\lambda^2}{2}\right) \quad \text{i.e., } \sigma \sim \text{subG}(1)$$

Therefore let $\lambda \to \frac{v_i}{n}\lambda$

$$\begin{aligned} \text{MGF}_{\sigma_i\sim\text{Rad}}\left(\frac{v_{j,i}}{n}\lambda\right) &\leq \exp\left(\left(\frac{v_i\lambda}{n}\right)^2\frac{1}{2}\right) = \exp\left(\frac{v_{j,i}^2\lambda^2}{2n^2}\right) \\ &= \exp\left(\frac{\lambda^2}{2n^2}\right) \quad \text{since } v_{j,i} \in \{-1,1\} \implies v_{j,i}^2 = 1 \end{aligned} \tag{13}$$

since v_i disappears from the weights, each term below now has identical MGF for i.i.d., σ_i:

$$
\begin{aligned}
\implies \text{MGF}_{\sum_{i=1}^{n}\sigma_i}\left(\frac{v_{j,i}}{n}\lambda\right) &\leq \exp\left(\frac{\lambda^2}{2n^2}\times n\right) \\
&= \exp\left(\frac{\lambda^2}{2n}\right) = \exp\left(\frac{1}{n}\frac{\lambda^2}{2}\right) \\
\implies t^2 &= \frac{1}{n} \\
\implies t &= \frac{1}{\sqrt{n}}
\end{aligned}
\tag{14}
$$

2.1.5 putting it all together

What is n in this setting? It's not the number of data point n, but instead it's the number of elements of the set: $|\mathbf{V}_n| = |\mathcal{H}_{Z_1,\ldots,Z_n}|$

using **Theorem(5)**:

$$
\begin{aligned}
\mathbb{E}_{\bar{\sigma}}\big[\max\{X_1,\ldots,X_n\}\big] &\leq t\sqrt{2\log(n)} \\
\implies \widehat{\text{Rad}}_S(\mathcal{H}) = \mathbb{E}_{\bar{\sigma}}\left[\max\left\{\Big(\sum_{i=1}^{n}\frac{\sigma_i v_i}{n}\Big)_{V_1},\ldots,\Big(\sum_{i=1}^{n}\frac{\sigma_i v_i}{n}\Big)_{V_{|V|}}\right\}\right] &\leq \frac{1}{\sqrt{n}}\sqrt{2\log(|\mathbf{V}_n|)} \\
&\leq \sqrt{\frac{2\log(|\mathbf{V}_n|)}{n}}
\end{aligned}
\tag{15}
$$

now we add the outer expectation $\mathbb{E}_{\bar{Z}}[\cdot]$ into, we have:

$$
\begin{aligned}
\text{Rad}_n(\mathcal{H}) &= \mathbb{E}_{\bar{Z}}\Big[\mathbb{E}_{\bar{\sigma}}\Big[\max_{V_j\in\mathcal{H}_{Z_1,\ldots,Z_n}}\Big\{\frac{\bar{\sigma}^\top V_j}{n}\Big\}_{j=1}^{|\mathbf{V}_n|}\;\Big|\;\bar{Z}\Big]\Big] \\
&\leq \mathbb{E}_{\bar{Z}}\Big[\sqrt{\frac{2\log(|\mathbf{V}_n|)}{n}}\Big] \\
&\leq \sqrt{\frac{2\log s(\mathcal{H},n)}{n}} \qquad s(\mathcal{H},n) = \max_{\bar{Z}}|\mathbf{V}_n|
\end{aligned}
\tag{16}
$$

3 Bounds on Expection of Maximum

Theorem 5 *Let $X_1, \dots, X_n$ be random variables. Suppose there exists $\sigma > 0$ s.t.:*

$$\mathbb{E}\big[\exp^{(\lambda X_i)}\big] \leq \exp^{\left(\frac{\lambda^2\sigma^2}{2}\right)} \quad \forall \lambda > 0 \tag{18}$$

then:

$$\mathbb{E}\big[\max\{X_1, \dots, X_n\}\big] \leq \sigma\sqrt{2\log(n)} \tag{19}$$

3.1 notes about Theorem(5)

3.1.1 looks like SubG!

note that the assumption is relaxed than the definition of $X_i \sim \text{SubG}(\sigma^2)$, as we need to have:

$$\mathbb{E}\big[\exp^{(\lambda X_i)}\big] \leq \exp^{\left(\frac{\lambda^2\sigma^2}{2}\right)} \quad \forall \lambda \in \mathbb{R}$$

therefore, if $X_i \sim \text{SubG}(\sigma^2)$ it is also suitable to use **Theorem (5)**

3.1.2 no i.i.d assumption on $\{X_i\}$

Theorem (5) has no i.i.d. assumption on $\{X_i\}$, otherwise one may no apply this to bound Eq.(12)

$$\mathbb{E}_{\bar{\sigma}}\Big[\max_{V_j \in \mathcal{H}_{Z_1,\dots,Z_n}} \Big\{\frac{\bar{\sigma}^\top V_j}{n}\Big\}_{j=1}^{|V|}\Big] \tag{20}$$

3.1.3 proof

first, let's wrap it around with: $\exp(\lambda\cdot)$, so we can use Jensen's inequlity to bring less than:

$$\begin{aligned}
&\exp\Big(\lambda\mathbb{E}\big[\max\{X_1, \dots, X_n\}\Big) \\
&\quad\leq \mathbb{E}\big[\exp^{\left(\lambda \max\{X_1,\dots,X_n\}\right)}\big] \\
&\quad= \mathbb{E}\big[\max\{\exp^{(\lambda X_1)}, \dots, \exp^{(\lambda X_n)}\}\big] \qquad \exp^{\lambda\max\{\cdot\}} = \max\{\exp^{\lambda(\cdot)}\} \quad \text{if } \lambda > 0 \\
&\quad\leq \mathbb{E}\Big[\sum_i^n \exp^{(\lambda X_i)}\Big] \quad \text{each term is non-negative, union bound, no iid assumption} \\
&\quad= \sum_i^n \mathbb{E}\Big[\exp^{(\lambda X_i)}\Big] \\
&\quad\leq n\exp^{\left(\frac{\lambda^2\sigma^2}{2}\right)} \quad \forall\lambda > 0 \quad \text{bring the bound assumption}
\end{aligned} \tag{21}$$

re-arrange terms to have only

$$\begin{aligned}
\exp\Big(\lambda\mathbb{E}\big[\max\{X_1, \dots, X_n\}\big]\Big) &\leq n\exp^{\left(\frac{\lambda^2\sigma^2}{2}\right)} \\
\lambda\mathbb{E}\big[\max\{X_1, \dots, X_n\}\big] &\leq \log(n) + \frac{\lambda^2\sigma^2}{2} \\
\mathbb{E}\big[\max\{X_1, \dots, X_n\}\big] &\leq \frac{\log(n)}{\lambda} + \frac{\lambda\sigma^2}{2}
\end{aligned} \tag{22}$$

since any λ works, we can just minimize $\frac{\log(n)}{\lambda} + \frac{\lambda\sigma^2}{2}$

$$\begin{aligned}
\implies \frac{\sigma^2}{2} &= \frac{\log(n)}{\lambda^2} \\
\implies \lambda^2 &= \frac{2\log(n)}{\sigma^2} \\
\implies \lambda &= \frac{\sqrt{2\log(n)}}{\sigma}
\end{aligned} \tag{23}$$

QUESTION should we must check $\lambda > 0$?
substitute back:

$$\begin{aligned}
\mathbb{E}\big[\max\{X_1, \ldots, X_n\}\big] &\leq \frac{\log(n)}{\lambda} + \frac{\lambda\sigma^2}{2} \\
&= \frac{\log(n)\sigma}{\sqrt{2\log(n)}} + \frac{\sqrt{2\log(n)}\sigma^2}{2\sigma} \\
&= \frac{\sqrt{\log(n)}\sigma}{\sqrt{2}} + \frac{\sqrt{2\log(n)}\sigma}{\sqrt{2}} \\
&= \sigma\sqrt{2\log(n)}
\end{aligned} \tag{24}$$

note that this is a "hard bound", meaning:

$$\Pr\Big(\mathbb{E}\big[\max\{X_1, \ldots, X_n\}\big] \leq \sqrt{2\log(n)}\sigma\Big) = 1 \tag{25}$$

4 Rademacher Complexity on linear models

now we extend to more specific models, such as Linear and Neural Networks

Theorem 6 *Let $\mathcal{H} = \{\mathbf{x} \to \boldsymbol{w}^\top \mathbf{x} : \|\boldsymbol{w}\|_2 \le B$, and assume $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\|\mathbf{x}\|^2 \le C^2\}$. Then:*

$$\widehat{Rad}_S(\mathcal{H}) \le \frac{B}{n}\sqrt{\sum_i \|\mathbf{x}_i\|_2^2} \tag{26}$$

and

$$Rad_n(\mathcal{H}) \le \frac{BC}{\sqrt{n}} \tag{27}$$

4.1 proof

$$\begin{aligned}
\widehat{\mathrm{Rad}}_S(\mathcal{H}) &= \mathbb{E}_\sigma \sup_{\|\boldsymbol{w}\|_2 \le B} \frac{1}{n}\sum_i \sigma_i \boldsymbol{w}^\top \mathbf{x}_i \quad \text{Empirical Rademacher complexity} \\
&= \frac{1}{n}\mathbb{E}_\sigma \sup_{\|\boldsymbol{w}\|_2 \le B} \boldsymbol{w}^\top \Big(\sum_i \sigma_i \mathbf{x}_i\Big) \\
&= \frac{B}{n}\mathbb{E}_\sigma \Big\|\sum_i \sigma_i \mathbf{x}_i\Big\|_2 \\
&\text{this is dual norm problem: } \|\mathbf{x}\|_* = \sup_{\|\boldsymbol{w}\|_2 \le 1} \boldsymbol{w}^\top \mathbf{x} \quad L_2 \text{ is self-norm}
\end{aligned} \tag{28}$$

4.1.1 a little detour: dual norm

QUESTION can you show L_2 is self-norm, i.e, why $\sup_{\|\boldsymbol{w}\|_2 \le 1} \boldsymbol{w}^\top \mathbf{x} = \|\mathbf{x}\|_2$?

QUESTION what is the dual norm of L_1?, i.e., what is $\sup_{\|\boldsymbol{w}\|_1 \le 1} \boldsymbol{w}^\top \mathbf{x}$?

QUESTION what is the dual norm of L_1?, i.e., what is $\sup_{\|\boldsymbol{w}\|_\infty \le 1} \boldsymbol{w}^\top \mathbf{x}$?

QUESTION A systematic answer using Holder's inequality? $\|\boldsymbol{w} \odot \mathbf{x}\|_1 \le \|\boldsymbol{w}\|_p \|\mathbf{x}\|_q \quad \frac{1}{p} + \frac{1}{q} = 1$

now we have:

$$
\begin{aligned}
\widehat{\text{Rad}}_S(\mathcal{H}) &= \frac{B}{n}\mathbb{E}_\sigma \underbrace{\Big\| \sum_i \sigma_i \mathbf{x}_i \Big\|_2}_{z} \\
&\equiv \frac{B}{n}\mathbb{E}_\sigma[z] \qquad \text{let } z = \Big\| \sum_i \sigma_i \mathbf{x}_i \Big\|_2 \\
&= \frac{B}{n}\mathbb{E}_\sigma[\sqrt{z^2}] \\
&\leq \frac{B}{n}\big(\mathbb{E}_\sigma[z^2]\big)^{\frac{1}{2}} \qquad \sqrt{t} \text{ is concave} \\
&= \frac{B}{n}\Big(\mathbb{E}_\sigma\Big[\Big\| \sum_i \sigma_i \mathbf{x}_i \Big\|_2^2\Big]\Big)^{\frac{1}{2}} \qquad \text{substitute back } z = \Big\| \sum_i \sigma_i \mathbf{x}_i \Big\|_2
\end{aligned}
\tag{31}
$$

$$
\begin{aligned}
\text{looking at:} \Big\| \sum_i \sigma_i \mathbf{x}_i \Big\|_2^2 &= \left\| \begin{matrix} \sum_{i=1}^n \sigma_i x_{i,1} \\ \vdots \\ \sum_{i=1}^n \sigma_i x_{i,d} \end{matrix} \right\|_2^2 \\
&= \sum_{k=1}^d \Big(\sum_{i=1}^n \sigma_i x_{i,k} \Big)^2 \\
&= \sum_{k=1}^d \Big(\sum_{i=1}^n \sum_{j=1}^n \sigma_i \sigma_j x_{i,k} x_{j,k} \Big) = \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^d \sigma_i \sigma_j x_{i,k} x_{j,k} \\
&= \sum_{i=1}^n \sum_{j=1}^n \sigma_i \sigma_j \sum_{k=1}^d x_{i,k} x_{j,k} \\
&= \sum_{i=1}^n \sum_{j=1}^n \sigma_i \sigma_j \mathbf{x}_i^\top \mathbf{x}_j
\end{aligned}
\tag{32}
$$

therefore looking at:

$$
\mathbb{E}\Big[\sum_{i=1}^n \sum_{j=1}^n \sigma_i \sigma_j \mathbf{x}_i^\top \mathbf{x}_j \Big] = \mathbb{E}\Big[\sigma_i^2 \mathbf{x}_i^\top \mathbf{x}_i + 2 \sum_{i=1}^n \sum_{j>i}^n \sigma_i \sigma_j \mathbf{x}_i^\top \mathbf{x}_j \Big]
\tag{33}
$$

substitute Eq.(31), we have:

$$
\begin{aligned}
\widehat{\text{Rad}}_S(\mathcal{H}) &\leq \frac{B}{n}\left(\mathbb{E}_\sigma \left[\sum_{i=1}^n \sigma_i^2 \mathbf{x}_i^\top \mathbf{x}_i + 2 \sum_{i=1}^n \sum_{j>i}^n \sigma_i \sigma_j \mathbf{x}_i^\top \mathbf{x}_j \right] \right)^{\frac{1}{2}} \\
&= \frac{B}{n}\left(\sum_{i=1}^n \mathbb{E}_\sigma[\sigma_i^2] \mathbf{x}_i^\top \mathbf{x}_i + 2 \sum_{i=1}^n \sum_{j>i}^n \mathbb{E}_\sigma[\sigma_i]\mathbb{E}_\sigma[\sigma_j] \mathbf{x}_i^\top \mathbf{x}_j \right)^{\frac{1}{2}} \\
&= \frac{B}{n}\left(\sum_{i=1}^n \mathbf{x}_i^\top \mathbf{x}_i \right)^{\frac{1}{2}} \qquad \text{as } \mathbb{E}_\sigma[\sigma_i^2] = 1 \quad \mathbb{E}_\sigma[\sigma_i] = 0 \\
&= \frac{B}{n}\sqrt{\sum_i \|\mathbf{x}_i\|_2^2}
\end{aligned}
\tag{34}
$$

$$
\begin{aligned}
\mathrm{Rad}_n(\mathcal{H}) &= \mathbb{E}_S\left[\widehat{\mathrm{Rad}}_S(\mathcal{H})\right] \\
&\leq \frac{B}{n}\mathbb{E}_S\Big[\sqrt{\sum_{i=1}^{n}\|\mathbf{x}_i\|_2^2}\Big] \\
&\leq \frac{B}{n}\sqrt{\mathbb{E}_S\big[\sum_{i=1}^{n}\|\mathbf{x}_i\|_2^2\big]} \qquad \sqrt{t}\ \text{ is concave} \\
&= \frac{B}{n}\sqrt{\sum_{i=1}^{n}\mathbb{E}_{\mathbf{x}_i}\big[\|\mathbf{x}_i\|_2^2\big]} \quad \text{swap} \sum \text{ and } \mathbb{E}[\cdot] \\
&\leq \frac{B}{n}\sqrt{C^2 n} \qquad \text{assumption } \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\|\mathbf{x}\|^2 \leq C^2 \\
&= \frac{BC}{\sqrt{n}}
\end{aligned}
\tag{35}
$$

5 neural networks

Theorem 8 *Let $\mathcal{H} = \{h_\theta : \|\boldsymbol{w}\|_2 \leq B' \text{ and } \|\mathbf{u}_i\| \leq B \quad \forall i\}$ Then:*

$$Rad_n(\mathcal{H}) \leq 2BB'C\sqrt{\frac{m}{n}} \tag{38}$$

where $h_\theta \equiv \boldsymbol{w}^\top \phi(\mathbf{U}\mathbf{x}_i)$

$\boldsymbol{w}$ is a vector and $\mathbf{U}$ is a matrix, and bound is place for each i^{th} row of $\mathbf{U}$, i.e., $\mathbf{u}_j$

5.1 proof

starting by proving $\widehat{\text{Rad}}_S(\mathcal{H})$ first:

$$\begin{aligned}
\widehat{\text{Rad}}_S(\mathcal{H}) &= \mathbb{E}_\sigma \sup_{\boldsymbol{w},U} \frac{1}{n}\sum_i \sigma_i \boldsymbol{w}^\top {\color{red}\phi(\mathbf{U}\mathbf{x}_i)} \quad \text{compared with linear model } \mathbf{x} \to {\color{red}\phi(\mathbf{U}\mathbf{x}_i)} \\
&= \mathbb{E}_\sigma \sup_{\boldsymbol{w},\mathbf{U}} \boldsymbol{w}^\top \left(\frac{1}{n}\sum_i \sigma_i \phi(\mathbf{U}\mathbf{x}_i)\right) \quad \text{not taking } \frac{1}{n} \text{ out for a reason (later)} \\
&\text{this is not dual norm problem before } \|\mathbf{x}\|_* = \sup_{\|\boldsymbol{w}\|_2 \leq 1} \boldsymbol{w}^\top \mathbf{x} \quad \text{since } \mathbf{x} \text{ also varies} \\
&= \mathbb{E}_\sigma \sup_{\boldsymbol{w},\mathbf{U}} \|\boldsymbol{w}\|_2 \left\|\frac{1}{n}\sum_i \sigma_i \phi(\mathbf{U}\mathbf{x}_i)\right\|_2
\end{aligned} \tag{39}$$

maximum occurs when $\boldsymbol{w}$ and $\sum_i \sigma_i \phi(\mathbf{U}\mathbf{x}_i)$ in the same direction:

$$\begin{aligned}
\mathbf{u}^\top \mathbf{v} &\leq \|\mathbf{u}\|\|\mathbf{v}\| \quad \text{Cauchy–Schwarz} \\
\implies \sup_{\mathbf{u},\mathbf{v}} \mathbf{u}^\top \mathbf{v} &= \|\mathbf{u}\| \; \|\mathbf{v}\| \quad \text{when } \mathbf{u}, \mathbf{v} \text{ in same direction}
\end{aligned} \tag{40}$$

One may think, we can just maximize $\sup_{\mathbf{U}} \|\sum_i \sigma_i \phi(\mathbf{U}\mathbf{x}_i)\|$. Condition on optimized $\mathbf{U}$, we can then orient $\boldsymbol{w}$ to maximize $\boldsymbol{w}$ (which gives B'):

$$\begin{aligned}
\widehat{\text{Rad}}_S(\mathcal{H}) &= B'\mathbb{E}_\sigma \sup_{\mathbf{U}} \left\|\frac{1}{n}\sum_i \sigma_i \phi(\mathbf{U}\mathbf{x}_i)\right\|_2 \quad \text{apply } \|\boldsymbol{w}\|_2 \leq B' \\
&= B'\mathbb{E}_\sigma \sup_{\|\mathbf{u}_j\| \leq B \; \forall j} \left\|\frac{1}{n}\sum_i \sigma_i \phi(\mathbf{U}\mathbf{x}_i)\right\|_2 \quad \text{apply } \|u_j\|_2 \leq B \\
&= B'\mathbb{E}_\sigma \underbrace{\sup_{\|\mathbf{u}_j\| \leq B} \left\| \begin{bmatrix} \frac{1}{n}\sum_i \sigma_i \phi(\mathbf{u}_{1,:}^\top \mathbf{x}_i) & \dots & \frac{1}{n}\sum_i \sigma_i \phi(\mathbf{u}_{m,:}^\top \mathbf{x}_i) \end{bmatrix}^\top \right\|_2}_{\sup_{v_1,\dots,v_m} \sqrt{\sum_{j=1}^m f(v_j)^2}}
\end{aligned} \tag{41}$$

$$\begin{aligned}
\sup_{v_1,\dots,v_m} \sqrt{\sum_{j=1}^m f(v_j)^2} &= \sqrt{\sum_{j=1}^m \sup_{v_j} f(v_j)^2} \quad \text{since each } v_j \text{ can be optimized independently} \\
&= \sqrt{m \sup_v f(v)^2} \quad \text{and in identical fashion} \\
&= \sqrt{m} \sup_v |f(v)|
\end{aligned} \tag{42}$$

substitute $f(v) = \sum_i \sigma_i \phi(\mathbf{u}_{j,:}\mathbf{x}_i)$ for any $j \in 1 \ldots m$, and let $\mathbf{u}$ be a particular $\mathbf{u}_j$:

$$\begin{aligned}
\widehat{\text{Rad}}_S(\mathcal{H}) &= B'\sqrt{m}\mathbb{E}_\sigma \sup_{\|\mathbf{u}\|_2 \le B} \left| \frac{1}{n} \sum_i \sigma_i \phi(\mathbf{u}^\top \mathbf{x}_i) \right| \\
&\le 2B'\sqrt{m}\mathbb{E}_\sigma \sup_{\|\mathbf{u}\|_2 \le B} \left| \frac{1}{n} \sum_i \sigma_i (\mathbf{u}^\top \mathbf{x}_i) \right| \quad \text{Talagrand Lemma [1]} \\
&= 2B'\sqrt{m} \underbrace{\mathbb{E}_\sigma \sup_{\|\mathbf{u}\|_2 \le B} \left(\frac{1}{n} \sum_i \sigma_i (\mathbf{u}\mathbf{x}_i) \right)}_{\widehat{\text{Rad}}_S(\mathcal{H})\left(\mathcal{H}=\{x \to \mathbf{u}^\top \mathbf{x} :\ \|\mathbf{u}\|_2 \le B)\}\right)} \quad \text{assume positive activation}
\end{aligned} \tag{43}$$

$$\begin{aligned}
\text{Rad}_n(\mathcal{H}) &= \mathbb{E}_S\left[\widehat{\text{Rad}}_S(\mathcal{H})\right] \\
&\le 2B'\sqrt{m}\mathbb{E}_S\left[\widehat{\text{Rad}}_S\left(\mathcal{H} = \{x \to \mathbf{u}^\top \mathbf{x} :\ \|\mathbf{u}\|_2 \le B)\}\right)\right] \\
&\le 2B'\sqrt{m}\frac{BC}{\sqrt{n}} \\
&= 2B'BC\sqrt{\frac{m}{n}}
\end{aligned} \tag{44}$$

6 homework

Read up the following:

1. general concept of PAC Bayesian
2. and to read

7 references

in this tutorial, I have paraphrased a number of existing courses and notes, I encourage people to see the original notes too.

1. `http://www.stat.cmu.edu/~larry/=sml/Concentration-of-Measure.pdf`
2. `https://web.stanford.edu/class/cs229t/scribe_notes/10_15_final.pdf`
3. various Wikipedia pages

References

[1] Peter L Bartlett and Shahar Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 463–482, 2002.