Chapter6. Empirical Risk Minimization: Abstract risk bounds and Rademacher averages

May 31, 2021

Last chapter

$\hat{f}_n = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \; \mathbb{E}_{P_n} \ell_f(z)$

Theorem 5.3

Empirical Risk Minimization (ERM) algorithm is a PAC algorithm if

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} P^n \left(\boxed{\sup_{f\in\mathcal{F}} \left| P_n\left(\ell_f\right) - P\left(\ell_f\right) \right|} \geq \varepsilon \right) = 0, \quad \forall \varepsilon > 0,$$

for every $\epsilon > 0$..

Notations

Agnostic (model-free) learning

- Sets $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{U}$
- A class $\mathcal{P}$ of probability distributions on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$
- A class $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathcal{U}$ (the hypothesis space)
- A loss function $\ell : \mathcal{Y} \times \mathcal{U} \to [0, 1]$

Notations

Agnostic (model-free) learning

- Sets $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{U}$
- A class $\mathcal{P}$ of probability distributions on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$
- A class $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathcal{U}$ (the hypothesis space)
- A loss function $\ell : \mathcal{Y} \times \mathcal{U} \to [0, 1]$

An abstract framework for ERM

- Set $\mathcal{Z}$
- A class $\mathcal{P}$ of probability distributions on $\mathcal{Z}$
- A class $\mathcal{F}$ of functions $f : \mathcal{Z} \to [0, 1]$ (induced losses)

$\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$

$z = (x, y) \quad g \in \mathcal{H}$

$\ell(g(x), y) = f(z)$

Notations

- The expected risk of any $f \in \mathcal{F}$:

$$L_P(f) \quad P(f) := \mathbf{E}_P f(Z)$$

- The minimum risk :

$$L_P^*(\mathcal{F}) := \inf_{f \in \mathcal{F}} P(f)$$

- ERM algorithm:

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} P_n(f) = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(Z_i)$$

$$P_n(f) = \mathbb{E}_{P_n} f(z) = L_{P_n}(f)$$

- Uniform deviation

$$\Delta_n(Z^n) := \|P_n - P\|_{\mathcal{F}} \equiv \underline{\sup_{f \in \mathcal{F}}} |P_n(f) - P(f)|$$

An abstract framework for ERM

- Set $\mathcal{Z}$
- A class $\mathcal{P}$ of probability distributions on $\mathcal{Z}$
- A class $\mathcal{F}$ of functions $f : \mathcal{Z} \to [0,1]$ (induced losses)

$\mathcal{F}$-seminorm $\|\cdot\|_{\mathcal{F}}$

For $P, P' \in \mathcal{P}$

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

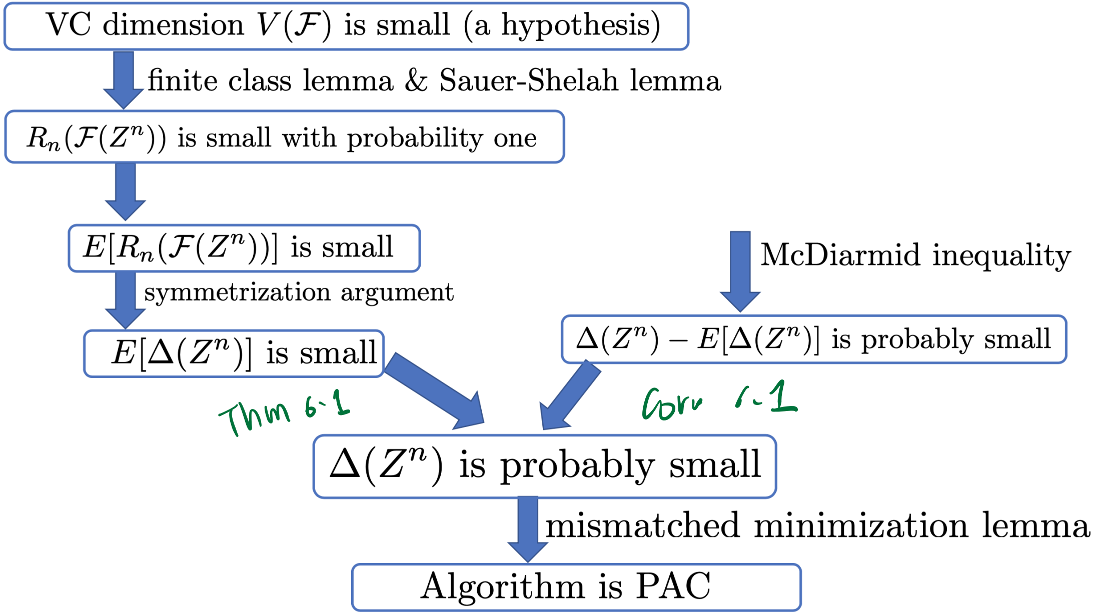
Why bounding the uniform deviation

Proposition 6.1

The generalization loss for a learning algorithm satisfies:

$$P\left(\hat{f}_n\right) \leq L^*(\mathcal{F}) + 2\Delta_n\left(Z^n\right) \text{(if algorithm is ERM)}$$

$$P\left(\widehat{f_n}\right) \leq P_n\left(\widehat{f_n}\right) + \Delta_n\left(Z^n\right) \text{ (for any algorithm).}$$


VC dimension $V(\mathcal{F})$ is small (a hypothesis)
finite class lemma & Sauer-Shelah lemma
$R_n(\mathcal{F}(Z^n))$ is small with probability one
$E[R_n(\mathcal{F}(Z^n))]$ is small
symmetrization argument
$E[\Delta(Z^n)]$ is small
McDiarmid inequality
$\Delta(Z^n) - E[\Delta(Z^n)]$ is probably small
Thm 6.1
Corr 6.1
$\Delta(Z^n)$ is probably small
mismatched minimization lemma
Algorithm is PAC

Bounding the uniform deviation: Rademacher averages

Theorem 6.1

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

$$\mathbf{E}\Delta_n\left(Z^n\right) \leq 2\mathbf{E}R_n\left(\mathcal{F}\left(Z^n\right)\right)$$

Bounding the uniform deviation: Rademacher averages

Theorem 6.1

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

$$\mathbf{E}\Delta_n(Z^n) \le 2\mathbf{E}R_n\left(\boxed{\mathcal{F}(Z^n)}\right)$$

$$\mathcal{F}(Z^n) \triangleq \{ (f(z_1), \dots, f(z_n)) : f \in \mathcal{F} \}.$$

Definition 6.1.

Let $\mathcal{A} \subset \mathbb{R}^n$ with $\mathcal{A}$ bounded. The Rademacher average of $\mathcal{A}$, denoted by $R_n(\mathcal{A})$, is defined by

$$R_n(\mathcal{A}) = \mathbf{E}\left[\sup_{a \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i a_i \right| \right]$$

where $\varepsilon_1, \dots, \varepsilon_n$ are independent Rademacher (i.e., ± 1 with equal probability) random variables.

Bounding the uniform deviation: Rademacher averages

Theorem 6.1

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

$$\mathbf{E}\Delta_n(Z^n) \leq 2\mathbf{E}R_n(\mathcal{F}(Z^n))$$

Proof:

$$\mathbb{E}\,\Delta_n(Z^n) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - P(f)\right|\right]$$

Fact 1: $\{Z_i\} \overset{\text{i.i.d.}}{\sim} P$, $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} f(Z_i)\right] = \mathbb{E}_P f(Z) = P(f)$

Fact 2: $\phi : \{ y(f) = \mathbb{E}_y f(Z) ; f \in \mathcal{F} \} \to \mathbb{R}$

$y \in \mathcal{P}$

Jason 1

$$y \mapsto \mathbb{E}\Big[\sup_{f\in\mathcal{F}} \Big| \frac{1}{n}\sum_{i=1}^{n} f(z_i) - y(f) \Big| \Big]$$

$$\Delta_n(z^n) = \phi\big(\overbrace{P(f)}^{P(\mathcal{F})}\big) = \phi\Big(\mathbb{E}_{\bar{z}_i}\Big[\frac{1}{n}\sum_{i=1}^{n} f(\bar{z}_i)\Big]\Big)$$

$$\leq \mathbb{E}_{\bar{z}_i}\Big[\phi\Big[\frac{1}{n}\sum_{i=1}^{n} f(\bar{z}_i)\Big]\Big)$$

$$= \mathbb{E}_{\bar{z}_i}\Big[\sup_{f\in\mathcal{F}} \Big|\frac{1}{n}\sum_{i=1}^{n} \underbrace{\big(f(z_i) - f(\bar{z}_i)\big)}\Big|\Big]$$

$t \in \mathbb{R}$.

$$M_1(t) = \mathbb{E}\big[e^{t(f(z_i) - f(\bar{z}_i))}\big]$$

$$= \mathbb{E}\big[e^{tf(z_i)}\big] \cdot \mathbb{E}\big[e^{-tf(\bar{z}_i)}\big]$$

$$M_2(t) = \mathbb{E}\big[e^{tf(\bar{z}_i)}\big] \cdot \mathbb{E}\big[e^{-tf(z_i)}\big]$$

$$\underbrace{f(z_i) - f(\bar{z}_i)}_{①} \overset{d}{=} \underbrace{f(\bar{z}_i) - f(z_i)}_{②} \overset{d}{=} \overbrace{\varepsilon_i\big(f(z_i) - f(\bar{z}_i)\big)}$$

$$\mathbb{E}_{z^n}\Delta_n(z^n) \leq \mathbb{E}_{z^n, \bar{z}_i, \varepsilon_i}\Big[\sup_{f\in\mathcal{F}} \Big|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i\big(f(z_i) - f(\bar{z}_i)\big)\Big|\Big]$$

$$\leq \mathbb{E}\left[\sup \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(z_i) \right| \right] + \mathbb{E}\left[\sup \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(z_i') \right| \right]$$

$$= 2 R_n(\mathcal{F}(z^n))$$

$\square$

Bounding the uniform deviation: Rademacher averages

Corollary 6.1

For any $P \in \mathcal{P}(\mathsf{Z})$ and any n, with probability at least $1-\delta$

$$\Rightarrow \quad P\left(\widehat{f}_n\right) \le L^*(\mathcal{F}) + 4\mathbf{E}R_n\left(\mathcal{F}\left(Z^n\right)\right) + \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}} \quad \text{(if ERM is used)}$$

$$P\left(\widehat{f}_n\right) \le P_n\left(\widehat{f}_n\right) + 2\mathbf{E}R_n\left(\mathcal{F}\left(Z^n\right)\right) + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}} \quad \text{(for any algorithm)}$$

$P(\hat{f}_n) \le P_n(\hat{f}_n) + \Delta_n(Z^n)$

Proof:

Coro 6.1: $P(\hat{f}_n) \le L^*(\mathcal{F}) + 2\Delta_n(Z^n) + 2\mathbb{E}\Delta_n(Z^n) - 2\mathbb{E}\Delta_n(Z^n)$

$\le L^*(\mathcal{F}) + 4R_n(\mathcal{F}(Z^n)) + 2\underbrace{\left(\Delta_n(Z^n) - \mathbb{E}\Delta_n(Z^n)\right)}$

w.p. $1-\delta$ $\le \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}$

McDiamid inequality :

$X^n = (X_1 \cdots X_n)$ n-tuple of indepedent r.v's

If a func $g : X^n \to \mathbb{R}$ has bound difference , i.e,

$$\sup_{x \in X} g(x_1 \ldots x_{i-1}, x, x_{i+1} \cdots x_n) - \inf g(x_1 \ldots x_{i-1}, x, x_{i+1} \cdots x_n) \le c_i$$

Then, for all $t > 0$

$$\mathbb{P}(g(X^n) - \mathbb{E}\, g(X^n) \ge t) \le \exp\left(- \frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right).$$

$$\Delta_n(Z^n) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \underbrace{\sum_{i=1}^{n} f(z_i)} - P(f) \right|$$

$$c_i = \frac{1}{n}$$

Let $\varepsilon = \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}$

$$\mathbb{P}\left(\Delta(Z^n) - \mathbb{E}\,\Delta_n(Z^n) \geq \varepsilon \right) \leq \exp\left(- \frac{2\varepsilon^2}{n \cdot \frac{1}{n^2}} \right) = \delta$$

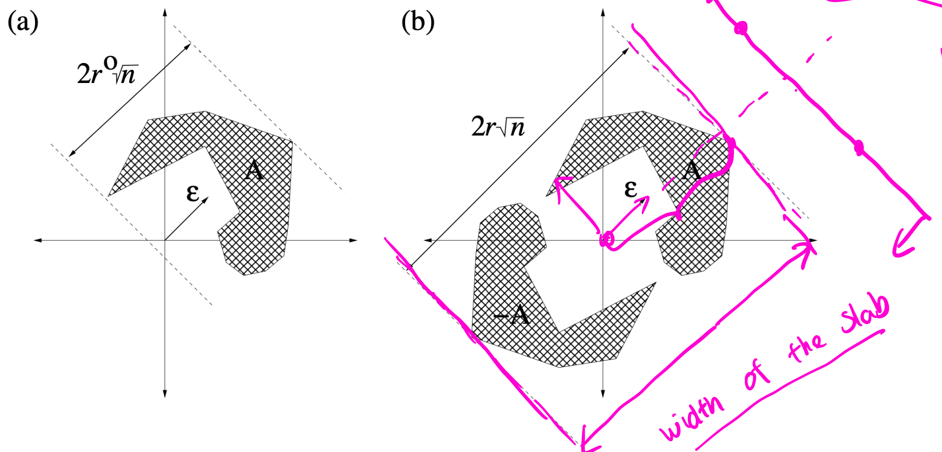
$\square$

Structural results for Rademacher averages

$\vec{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)$ $\quad \frac{1}{\sqrt{n}}\vec{\varepsilon}$ has length 1.

$$R_n(\mathcal{A}) := \frac{1}{n}\mathbf{E}_{\varepsilon^n}\left[\sup_{a\in\mathcal{A}}\left|\boxed{\sum_{i=1}^{n}\varepsilon_i a_i}\right|\right] = \frac{1}{\sqrt{n}}\mathbb{E}_{\varepsilon^n}\left[\sup_{\vec{a}\in\mathcal{A}}\left|\langle \frac{\vec{\varepsilon}}{\sqrt{n}}, \vec{a}\rangle\right|\right]$$

$$R_n^{\circ}(\mathcal{A}) := \frac{1}{n}\mathbf{E}_{\varepsilon^n}\left[\sup_{a\in\mathcal{A}}\sum_{i=1}^{n}\varepsilon_i a_i\right] = \mathbb{E}_{\varepsilon^n}\left[\underbrace{\frac{1}{\sqrt{n}}\sup_{\vec{a}\in\mathcal{A}\cup-\mathcal{A}}\langle \frac{\vec{\varepsilon}}{\sqrt{n}}, \vec{a}\rangle}_{r(\vec{\varepsilon})}\right]$$



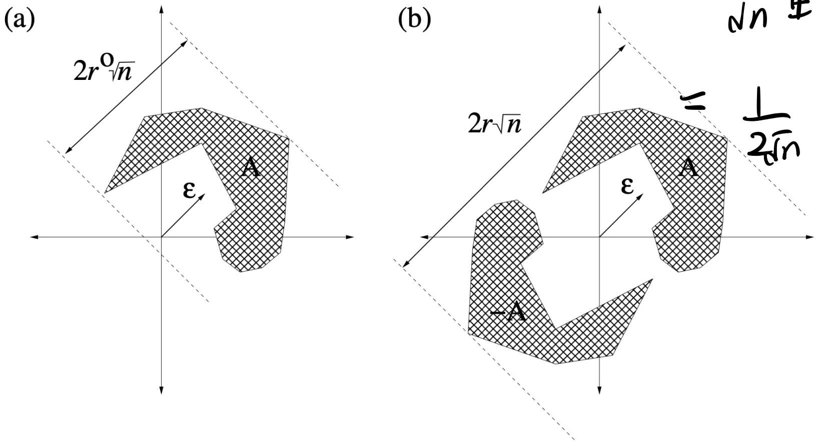
Structural results for Rademacher averages

$$R_n(\mathcal{A}) := \frac{1}{n}\mathbf{E}_{\varepsilon^n}\left[\sup_{a\in\mathcal{A}}\left|\sum_{i=1}^{n}\varepsilon_i a_i\right|\right]$$

$$R_n^{\circ}(\mathcal{A}) := \frac{1}{n}\mathbf{E}_{\varepsilon^n}\left[\sup_{a\in\mathcal{A}}\sum_{i=1}^{n}\varepsilon_i a_i\right] = \frac{1}{\sqrt{n}}\mathbb{E}\left[\sup_{a\in A}\langle \frac{\vec{\varepsilon}}{\sqrt{n}}, \vec{a}\rangle\right]$$

$$= \frac{1}{\sqrt{n}}\mathbb{E}\left[\inf_{a\in A} -\langle \frac{\vec{\varepsilon}}{\sqrt{n}}, \vec{a}\rangle\right]$$

$$= \frac{1}{2\sqrt{n}}\mathbb{E}\left[\sup_{a\in A}\langle \frac{\vec{\varepsilon}}{\sqrt{n}}, \vec{a}\rangle - \inf_{a\in A}\langle \frac{\vec{\varepsilon}}{\sqrt{n}}, \vec{a}\rangle\right]$$



Structural results for Rademacher averages

Basic properties of Rademacher averages

(1) $R_n^\circ(\mathcal{A}) \leq R_n(\mathcal{A}) = R_n^\circ(\mathcal{A} \cup -\mathcal{A})$

$\mathcal{A} + \vec{v} \triangleq \{ \vec{a} + \vec{v} : a \in \mathcal{A} \}$

(2) $R_n^\circ(\mathcal{A}) = R_n(\mathcal{A})$ if $\mathcal{A} = -\mathcal{A}$

(3) $R_n^\circ(\mathcal{A} + v) = R_n^\circ(\mathcal{A})$ for any $v \in \mathbb{R}^n$

(4) $R_n(\mathcal{A} \cup \mathcal{B}) \leq R_n(\mathcal{A}) + R_n(\mathcal{B})$

(5) $R_n^\circ(\mathcal{A} + \mathcal{B}) = R_n^\circ(\mathcal{A}) + R_n^\circ(\mathcal{B})$

(6) $R_n(c\mathcal{A}) = |c| R_n(\mathcal{A})$

(7) $R_n(\mathcal{A}) = R_n(\mathrm{conv}(\mathcal{A}))$

$\sum c_i \vec{a}_i$

(8) $R_n(\mathcal{A}) = R_n(\underline{\mathrm{absconv}}(\mathcal{A}))$

$\sum |c_i| = 1$

Lemma 6.1 (Finite class lemma).

If $\mathcal{A} = \{a^{(1)}, \dots, a^{(N)}\} \subset \mathbb{R}^n$ is a finite set with $\|a^{(j)}\| \leq L$ for all $j = 1, \dots, N$ and $N \geq 2$, then

$\mathcal{A} = \mathcal{F}(Z^n) = \{(f(Z_1) \cdots f(Z_n)) : f \in \mathcal{F}\}$

$$R_n(\mathcal{A}) \leq \frac{2L\sqrt{\log N}}{n}$$

Proof:

def 2.1 A r.v. X is subgaussian with scale parameter ν if X has finite mean & $\mathbb{E}[e^{s(X-\mathbb{E}X)}] \leq e^{\frac{s^2\nu^2}{2}}$ for $s \in \mathbb{R}$.

$N(0, \nu^2)$

Hoffding's lemma (Lemm 2.1)

Let X r.v $[b,c]$, then $\mathbb{E}[e^{s(X-\mathbb{E}X)}] \leq e^{s^2 \frac{(c-b)^2}{8}}$

- Any $X \in [b,c]$, X is subgaussian with scale parameter

$$v = \frac{c-b}{2}.$$

- If $S = \sum_{i=1}^{n} X_i$ where X_i are independent subgaussian with scale parameter v_i.

Then S is subgaussian, $v^2 = \sum_{i=1}^{n} v_i^2$

$$R_n(A) = \mathbb{E} \sup_{a \in A} \frac{1}{n} \sum_{i=1}^{n} |\varepsilon_i a_i|$$

$$\left[-\frac{a_i}{n}, \frac{a_i}{n} \right]$$

$$= \mathbb{E} \sup_{a \in A \cup -A} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i a_i$$

$$\underbrace{Y_k = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i a_i^{(k)}}_{v^2 = \sum_{i=1}^{n} \frac{a_i^2}{n^2} \leq \frac{L^2}{n^2}}$$

$$= \mathbb{E} \max \{ Y_1, -Y_1 \cdots Y_N, -Y_N \}.$$

want to show $\leq \dfrac{2L\sqrt{\log N}}{n}$.

<u>Maximal lemma for subgaussian r.v</u> (Lemma 2.3)

Suppose $X_1 \dots X_n$, $\mathbb{E}X_i = 0$ with scale parameter v.

Then $\mathbb{E}[\max_{i \in [n]} X_i] \le v\sqrt{2\log n}$

$$v = \frac{L}{n}$$

$$R(A) \le \frac{L}{n} \cdot \sqrt{2\log 2N} \le \frac{2L}{n}\sqrt{\log N}.$$

$\square$.

Structural results for Rademacher averages

Proposition 6.2 (Contraction principles for Rademacher averages).

If $\mathcal{A}$ is a bounded subset of $\mathbb{R}^n$ and for $i \in [n], \varphi_i : \mathbb{R} \to \mathbb{R}$ is an M -Lipschitz continuous function, then $R_n^{\circ}(\varphi \circ \mathcal{A}) \leq M R_n^{\circ}(\mathcal{A})$. Furthermore, if $\varphi_i(0) = 0$ for all i (i.e., $\varphi(\mathbf{0}) = \mathbf{0}$) then $R_n(\varphi \circ \mathcal{A}) \leq 2M R_n(\mathcal{A})$

Proof:

def: $\varphi \circ A$

Let $\varphi_i : \mathbb{R} \to \mathbb{R}$, $\varphi \circ \vec{v} \triangleq (\varphi_1(v_1), \varphi_2(v_2) \cdots \varphi_n(v_n))$.

$\varphi \circ A = \{ \varphi \circ \vec{v} : \vec{v} \in A \}$.

def: φ is M-Lipschitz continuous if for all x_1, x_2

$|\varphi(x_1) - \varphi(x_2)| \leq M |x_1 - x_2|$.

$$\varphi \circ \vec{v} = \varphi_1 \circ \varphi_2 \cdots \varphi_n \circ \vec{v}$$

WTS: $R_n^{\circ}(\varphi_i \circ A) \leq R_n^{\circ}(A)$ $\qquad \varphi_i = \varphi_1$

$$R_n^{\circ}(A) = \frac{1}{n}\mathbb{E}\left[\sup_{a \in A} \sum_{i=1}^{n} \varepsilon_i a_i\right]$$

$$= \frac{1}{n}\mathbb{E}\left[\sup_{a \in A}\left(a_1 + \sum_{i=2}^{n} \varepsilon_i a_i\right)\frac{1}{2}\right.$$

$$\left.+ \sup_{a' \in A}\left(-a_1' + \sum_{i=2}^{n} \varepsilon_i a_i'\right)\frac{1}{2}\right]$$

$$= \frac{1}{2n}\mathbb{E}\left[\sup_{a, a' \in A} \left|\underbrace{a_1 - a_1'}\right| + \sum_{i=2}^{n} \varepsilon_i a_i + \sum_{i=2}^{n} \varepsilon_i a_i'\right]$$

$$Rn^{\circ}(\varphi_1 \circ A) = \frac{1}{2n}\mathbb{E}\Big[\sup_{a_1, a_1' \in A} \underbrace{|\varphi(a_1) - \varphi(a_1')|} + \sum_{i=2}^{n} \varepsilon_i a_i + \sum_{i=2}^{n} \varepsilon_i a_i'\Big]$$

$$Rn^{\circ}(\varphi_1 \circ A) \le Rn^{\circ}(A).$$

$$Rn(\varphi \circ A) = Rn^{\circ}\big((\varphi \circ A) \cup (-\varphi \circ A)\big)$$

$$= Rn^{\circ}\big((\varphi \circ A) \cup (-\varphi \circ A) \cup \{\vec{0}\}\big).$$

$Rn^{\circ}(A \cup B)$
if $\vec{0} \in A, B$.
then $Rn^{\circ}(A) + Rn^{\circ}(B) \geq Rn^{\circ}(A \cup B)$

$$\le Rn^{\circ}\big((\varphi \circ A) \cup \{\vec{0}\}\big) + Rn^{\circ}\big(-\varphi \circ A \cup \{\vec{0}\}\big).$$

$$= Rn^{\circ}\big(\varphi \circ (A \cup \{\vec{0}\})\big) + Rn^{\circ}\big(-\varphi(A \cup \{\vec{0}\})\big)$$

$$\le R_n^\circ(A \cup \{\vec{0}\}) + R_n^\circ(-A\{\vec{0}\}).$$

$R_n^\circ(aA) = |a| R_n^\circ(A)$

$$= 2R_n^\circ(\underline{A \cup \{\vec{0}\}})$$

$$\le 2R_n^\circ(\underline{A \cup -A})$$

$$= 2R_n(A).$$